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

Risco de Crédito de Municípios Brasileiros: uma abordagem utilizando métodos computacionais

Werley Antônio Mendonça Machado

Dissertação apresentada como requisito parcial para conclusão do

Mestrado Profissional em Computação Aplicada

Orientador

Prof. Dr. Edgard Costa Oliveira

Coorientadora

Prof.ª Dr.ª Ana Carla Bittencourt Reis

Brasília

2021

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

Risco de Crédito de Municípios Brasileiros: uma abordagem utilizando métodos computacionais

Werley Antônio Mendonça Machado

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

Prof. Dr. Edgard Costa Oliveira (Orientador)
EPR/UnB

Prof. Dr. João Carlos Félix Souza
CIC/UnB e EPR/UnB

Prof. Dr. Gustavo José de Guimarães e Souza
Ministério da Economia

Prof.ª Dr.ª Rafaela Mousinho Guidi
Banco do Brasil S.A.

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

Brasília, 03 de novembro de 2021

Ficha catalográfica elaborada automaticamente,

com os dados fornecidos pelo autor

MM149r

Machado, Werley Antônio Mendonça
Risco de Crédito de Municípios Brasileiros: Uma abordagem utilizando métodos computacionais / Werley Antônio Mendonça Machado; orientador Edgard Costa Oliveira; co-orientador Ana Carla Bittencourt Reis. -- Brasília, 2021.
137 p.

Dissertação (Mestrado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2021.

1. mineração de dados. 2. modelo de risco de crédito. 3. Probabilidade de default (PD). 4. risco setor público municipal. I. Oliveira, Edgard Costa, orient. II. Reis, Ana Carla Bittencourt, co-orient. III. Título.

Dedicatória

Dedico este trabalho aos meus amados pais, Nilton Vieira Machado (*in memorian*) e Wanda de Mendonça Machado (*in memorian*), que se entregaram e trabalharam de corpo e alma, com muito amor, na construção de um futuro melhor para os seus filhos.

Agradecimentos

Agradeço a Deus, que tudo fez e faz, por mim e por todos que me ajudaram a chegar até aqui. Agradeço à minha amada esposa Karine e aos meus amados filhos Vitor, Luísa e Maria Clara por me apoiarem, incentivarem, erguerem e reerguerem ao longo dessa caminhada em meio às dificuldades e turbulências vividas nesse período de estudo e aprendizado, sou eternamente grato por estarmos juntos na jornada da vida. Agradeço aos meus amados irmãos e amigos por me apoiarem e por entenderem minha ausência em vários momentos nos últimos anos.

Agradeço imensamente ao prof. Dr. Edgard Costa Oliveira cuja presença, contribuições e apoio foram fundamentais para que eu me mantivesse motivado a continuar em frente. Agradeço à prof.ª Dr.ª Ana Carla Bittencourt Reis pelas contribuições no trabalho e pelas excelentes aulas que me inspiraram na construção da pesquisa. Agradeço a todos os meus professores do Programa de Pós-graduação em Computação Aplicada (PPCA) e aos meus colegas de turma pelo aprendizado conjunto e trocas de experiências únicas, que levo comigo para a vida toda.

Agradeço de forma especial aos meus amigos que em incentivaram, me apoiaram e me acompanharam nessa jornada: Tiago Eny, Gustavo, João Flávio, Marinna, Bernardo, João Vicente, João Paulo, Eduardo Barbosa, André Baptista, Karen, Mietto, Claudia Ohtoshi, Rômulo. Vocês são show de bola!

Por fim, agradeço à empresa que trabalho e aos meus colegas de convívio pela oportunidade concedida em trocar horas de trabalho por investimento em estudo e conhecimento, sem esse apoio não seria possível a materialização desse sonho.

Resumo

A crise de 2008, também conhecida como "Subprime", afetou negativamente não só o sistema financeiro mundial, como também a economia produtiva dos países. Muitas corporações e empresas tiveram sua saúde financeira afetada, gerando uma onda voltada ao fortalecimento da gestão de riscos nas organizações, particularmente nas instituições financeiras. A crise também impactou fortemente os governos, central e local (municípios) que, de maneira geral, tiveram suas políticas fiscais influenciadas por altos níveis de dívida pública e de inadimplência. Esse fenômeno tem captado a atenção de pesquisadores e gestores de instituições públicas. O objetivo desta pesquisa é propor um modelo de decisão para classificar de forma massificada os municípios brasileiros quanto ao risco de crédito, observando as características dos entes públicos, tais como: nível de gestão pública, informações econômico-financeiras, características sociodemográficas e desenvolvimento social. O processo de desenvolvimento do modelo se baseou na norma ISO 31000 e no modelo de processo CRISP-DM, tido como o padrão em projetos de mineração de dados. Os resultados apresentados, mostraram que modelos desenvolvidos utilizando informações de dados abertos e métodos computacionais de *machine learning* (*XGBoost*) e métodos estatísticos tradicionais (Regressão Logística) são alternativas viáveis e com bons resultados, quanto à performance e à classificação e mensuração da probabilidade associada ao risco de crédito dos municípios. Apesar do grande avanço tecnológico e maior disponibilidade de informações nos últimos anos, este trabalho contribuiu para a redução do gap científico em relação a abordagens que utilizam esse tipo de técnica para o objetivo de estimar a inadimplência municipal.

Palavras-chave: mineração de dados, modelo de risco de crédito, Probabilidade de default (PD), risco setor público municipal.

Abstract

The 2008 crisis, also known as "Subprime", negatively affected not only the world financial system, but also the productive economy of countries. Many corporations and companies had their financial health affected, generating a wave aimed at strengthening risk management in organizations, particularly in financial institutions. The crisis also had a strong impact on central and local governments (municipalities), which, in general, had their fiscal policies influenced by high levels of public debt and defaults. This phenomenon has attracted the attention of researchers and managers of public institutions. The objective of this research is to propose a decision model to massively classify Brazilian municipalities in terms of credit risk, observing the characteristics of public entities, such as: level of public management, economic and financial information, sociodemographic characteristics and social development. The model development process was based on the ISO 31000 standard and the CRISP-DM process model, considered the standard in data mining projects. The results presented showed that models developed using information from open data and computational methods of machine learning (XGBoost) and traditional statistics (Logistic Regression) are viable alternatives with good results, regarding performance and classification and measurement of probability associated with the credit risk of the municipalities. Despite the great technological advances and greater availability of information in recent years, this work has contributed to reducing the scientific gap in relation to approaches that use this type of technique for the purpose of estimating municipal default.

Keywords: Data Mining, Credit Risk Model, Probability of Default (PD), Municipal Public Sector Risk.

Sumário

Lista de Tabelas

Lista de Gráficos

Lista de Figuras

Lista de Quadros

Lista de Abreviaturas e Siglas

ADASYN *Adaptive Synthetic Sampling*

AUC Área sob a curva ROC

Bacen Banco Central do Brasil

BIS Banco Internacional de Compensações

CAPAG Capacidade de Pagamento (informação da Secretaria do Tesouro Nacional)

CBSB Comitê de Basileia de Supervisão Bancária

CF Constituição Federal

CGU Controladoria Geral da União

CMN Conselho Monetário Nacional

CNN *Condensed Nearest Neighbor Rule*

CPM Credit Portfolio Management

CRISP-DM *CRoss Industry Standard Process for Data Mining*

EAD *Exposure at Default*

ECL Perdas Esperadas de Crédito

EUA Estados Unidos da América

FIRJAN Federação das Indústrias do Estado do Rio de Janeiro

FPR Taxa de falsos positivos

GIR Gerenciamento Integrado de Riscos e Capital

IACPM *International Association of Credit Portfolio Management*

IASB *International Accounting Standards Board*

IBGE Instituto Brasileiro de Geografia e Estatística

ICAAP *Internal Capital Adequacy Assessment Process*

IDHM Índice de Desenvolvimento Humano Municipal

IF Instituição Financeira

IFGF Índice Firjan de Gestão Fiscal

IFRS *International Financial Reporting Standards*

INSS Instituto Nacional do Seguro Social

IPEA Instituo de Pesquisa Econômica Aplicada

IRB *Internal Rating Based*

ISP Indicador de Situação Previdenciária

IVS Índice de Vulnerabilidade Social

KNN *K-nearest neghbors*

KS *Kolmogorov-Smirnov*

LRF Lei de Responsabilidade Fiscal

LGD *Loss Given Default*

NCR *Neighborhood Cleaning Rule*

NPL Ativos Considerados Problemáticos

OECD Organização para Cooperação e Desenvolvimento Econômico

OSS *One-Sided Selection*

PD Probabilidade de *Default*

PIB Produto Interno Bruto

PNAD Pesquisa Nacional por Amostra de Domicílios

PwC *Price-Waterhouse Coopers*

RAF *Risk Appetite Framework*

RCL Receita Corrente Líquida

RGPS Regime Geral da Previdência Social

RPPS Regime Próprio de Previdência Social

SFN Sistema Financeiro Nacional

Siconfi Sistema de Informações Contábeis e Fiscais do Setor Público Brasileiro

SMOTE *Synthetic Minority Oversampling Technique*

SVM *Suport Vetor Machine*

TEMAC Teoria do Enfoque Meta-Analítico Consolidado

TCU Tribunal de Contas da União

TPR Taxa de verdadeiros positivos

WoS *Web of Science*

XGBoost *Extreme Gradient Boosting*

1. Introdução

1.1. Contextualização

Em 2008, ocorreu a maior crise do sistema financeiro desde a "Grande depressão" de 1929. Essa crise, também conhecida como "*Subprime*", afetou negativamente não só o sistema financeiro mundial, como também a economia produtiva dos países, muitas corporações e empresas tiveram sua saúde financeira afetada. O valor perdido pelas empresas no mundo gerado por essa crise foi de US$ 14,5 trilhões [1], e nela o governo dos Estados Unidos a América (EUA) gastou bilhões em programas de Ativos Problemáticos para auxiliar as empresas em delicada situação econômico-financeira. Uma das causas para os pedidos de recuperação ou de falências das corporações era a fragilidade da gestão dos riscos aos quais estavam expostas [2].

Particularmente os governos dos países, de maneira geral tiveram suas políticas fiscais influenciadas por conta de altos níveis de dívida pública e inadimplência. Como consequência, ao redor do mundo os governos locais (municípios) também entraram em crise, e esse fenômeno tem captado a atenção de pesquisadores e gestores de instituições públicas com o objetivo de contribuir para a continuidade e sustentabilidade financeira dos serviços públicos [3]. Após a última crise financeira, as preocupações com as dívidas contraídas pelos municípios aumentaram significativamente. As dívidas municipais são particularmente preocupantes porque afetam não só o dia a dia dos cidadãos, mas também as empresas privadas locais, que dependem de decisões públicas.

Posteriormente à crise, o segmento mais afetado, o setor financeiro, vivenciou transformações voltadas a proporcionar maior segurança e resiliência às Instituições Financeiras (IF) e, por conseguinte, ao sistema financeiro dos países. O Banco Internacional de Compensações (BIS), responsável por estabelecer padrões internacionais de prudência financeira, por meio do Comitê de Basileia de Supervisão Bancária (CBSB), publicou, em dezembro de 2010 e uma versão revisada em junho de 2011, o marco regulatório conhecido como Basileia III [4], para orientar os bancos centrais dos países e as instituições financeiras a estabelecerem mecanismos de gestão de riscos e capital complementares aos padrões existentes, na versão anterior de Basileia II [5]. No Brasil, esses requisitos foram recepcionados e materializados na Resolução do Conselho Monetário Nacional (CMN) nº 4.557, de 23 de fevereiro de 2017 [6], também conhecida como Gerenciamento Integrado de Riscos e Capital (GIR), que dispõe sobre a estrutura de gerenciamento de riscos e a estrutura de gerenciamento de capital para fazer frente aos riscos que as IFs estão expostas.

Mesmo sendo um setor altamente regulado, novos negócios e instrumentos financeiros são criados para atender necessidades do mercado. Esse dinamismo, quando não acompanhado pelos supervisores do sistema financeiro, pode ser percebido como potencial gerador de crises, como a observada em 2008[7]. A resposta dos reguladores dos países para a crise do "*Subprime*" foi elevar o nível de exigência para as instituições financeiras com base nos padrões do CBSB. No pós-crise, observa-se um movimento crescente da indústria financeira no sentido de buscar soluções para melhorar o processo de gestão de risco e capital das IFs, de modo a reduzir os impactos da volatilidade dos ativos, dos mercados e dos segmentos da economia nos resultados e na rentabilidade dos bancos. Ainda sob o reflexo da crise, várias iniciativas relacionadas ao fortalecimento da governança e da gestão dos riscos nas organizações foram criadas [1].

Nesse contexto, faz-se necessário que as IFs invistam na melhoria dos processos voltados à gestão e desenvolvimento dos modelos para mensurar os riscos, contemplando todos os segmentos negociais e público atendidos, dentre eles os entes públicos como os municípios. Não somente para atendimento de exigências regulatórias, mas sobretudo para que as IFs tenham à disposição instrumentos robustos e acurados, voltados para o atendimento das necessidades de gestão dos riscos e negócios.

1.1.1. Características gerais dos municípios brasileiros

A República Federativa do Brasil é formada pela união de entes públicos, representados por 26 estados, 5.570 municípios e o Distrito Federal. Cada estado possui uma capital, uma cidade localizada em um município representativo do estado. As capitais além da importância como centro administrativo são o local onde está a representação dos poderes legislativo, executivo e judiciário dos estados [8].

As fontes de receitas dos municípios estão consagradas na Constituição Federal (CF) [8], e em Leis Complementares, Leis e Resoluções:

a) Receitas próprias do tipo "tributos", conforme art. 145 da CF:

I. *Impostos (art. 156 da CF): (i) propriedade predial e territorial urbana; (ii) transmissão inter vivos, a qualquer título, por ato oneroso, de bens imóveis, por natureza ou acessão física, e de direitos reais sobre imóveis, exceto os de garantia, bem como cessão de direitos a sua aquisição; (iii) serviços de qualquer natureza, não estando relacionados às operações relativas à circulação de mercadorias e sobre prestações de serviços de transporte interestadual e intermunicipal e de comunicação, ainda que as operações e as prestações se iniciem no exterior (não compreendidos no art. 155, II, da CF);*

II. *Taxas, em razão do exercício do poder de polícia ou pela utilização, efetiva ou potencial, de serviços públicos específicos e divisíveis, prestados ao contribuinte ou postos a sua disposição;*

III. *Contribuição de Melhoria, decorrente de obras públicas.*

b) Recursos provenientes de transferências constitucionais ou repartição de receitas, cuja fonte são os impostos da União e dos Estados, conforme art. 158 da CF;

c) Compensação financeira (royalties), pela exploração de petróleo ou gás natural, de recursos hídricos para fins de geração de energia elétrica e de outros recursos minerais no respectivo território, plataforma continental, mar territorial ou zona econômica exclusiva, conforme art. 20, §1º, da CF;

d) Patrimonial, pela exploração econômica do patrimônio público do município (bens móveis e imóveis), mediante aplicações financeiras, venda de bens móveis e imóveis, aluguéis;

e) De serviços, com a cobrança de tarifas sobre o transporte coletivo, mercados, feiras, cemitérios entre outros;

f) Outras receitas originadas de multas e outras penalidades administrativas (códigos de posturas, obras e outros regulamentos municipais, a atualização monetária e a cobrança da dívida ativa);

g) De Operações de Crédito: se enquadram nesse tipo as operações realizadas junto a instituições financeiras, como meio para suprir necessidade de recursos de caixa ou para cobrir despesas, cuja arrecadação orçamentária (tributos) não foi suficiente. Municípios podem contratar operações de crédito com instituições financeiras nacionais ou internacionais, nos termos do art. 32 da Lei Complementar nº 101, de 04 de maio de 2000, chamada Lei de Responsabilidade Fiscal (LRF) [9] e das Resoluções do Senado Federal 40/2001 [10], 43/2001 [11]e 48/2007 [12].

A crise econômica de 2015 e 2016 no Brasil gerou problemas na capacidade de prestação de serviços públicos pelos municípios, dada a escassez de recursos para executá-los. Na busca de outras fontes, dada a redução de receitas sobretudo das oriundas de transferências constitucionais, os municípios, com destaque para as capitais, têm recorrido a operações de crédito para fazer frente às suas necessidades de financiamento. Em 2019, a participação dos recursos originados por empréstimos e financiamentos junto a instituições financeiras, considerando 106 municípios selecionados pelo estudo "Multi Cidades - Finanças dos Municípios do Brasil", 2021 [13], representaram em média 37,9%

da fonte de recursos dos investimentos realizados pelos municípios. As operações de crédito a cada ano têm se tornado cada vez mais importantes no financiamento dos investimentos dessas cidades. A Figura 1.1 a seguir destaca a trajetória do endividamento das capitais e de grandes cidades entre 2006 e 2019.

Fonte: FNP [13]

***Figura 1.1**: Participação das operações de crédito em grandes municípios*

Observa-se, de 2010 aos anos mais recentes, uma trajetória crescente do endividamento, com a forte perda de capacidade de geração de recursos próprios. Particularmente, a partir de 2014, em decorrência da crise econômica, as prefeituras recorreram mais ao mercado de crédito. Existe, portanto, uma tendência de elevação nos gastos com juros e amortizações de dívidas no médio e no longo prazo devido à maturação das operações executadas no período mais recente. Segundo Relatório "Multi Cidades - Finanças dos Municípios do Brasil", de 2021[13], esse impacto será mais forte nos municípios maiores, pois são eles os grandes tomadores de crédito.

Mesmo o Brasil não tendo se recuperado totalmente da crise política e econômica do biênio 2015-2016, a crise mundial gerada pela pandemia do novo coronavírus, iniciada em 2020, colocou o Brasil em um cenário delicado. As ações mais que necessárias para conter os efeitos da pandemia na saúde dos brasileiros, resultaram, conforme análise do Banco Mundial [14], em três choques na economia, gerando a maior recessão já registrada no Brasil: (i) um choque externo, incluindo retração na demanda e queda de preços externos; (ii) um choque interno, com retração na demanda e na oferta gerada pelas restrições governamentais para evitar maior propagação de contágio da população; (iii) forte queda na demanda mundial de petróleo, provocando redução nos preços pela metade, com impacto negativo nas exportações brasileiras, tendo em vista que o Brasil é exportador líquido de petróleo.

Já no contexto dos entes públicos, segundo pesquisa IBOPE, de outubro de 2020 [15], o impacto causado pela pandemia do novo coronavírus nas contas públicas dos

municípios deve ser "alto" ou "muito alto", conforme opinião de sete em cada dez municípios participantes da pesquisa. Observou-se também que em grande parte dos municípios as políticas públicas (programas, medidas ou ações) previstas para promoção do desenvolvimento foram muito afetadas, particularmente as relacionadas à educação e geração de empregos.

1.2. Definição do problema de pesquisa

As IFs, nas ações de busca de novos negócios, atuam junto a diversos públicos e segmentos, como por exemplo, pessoas físicas, empresas de Atacado ou Varejo, cooperativas, sociedades sem fins lucrativos, produtores rurais, e também entes públicos (estados e municípios). Esses últimos são considerados clientes diferenciados e demandam soluções em serviços e necessidades de crédito, como financiamentos voltados para a modernização da administração tributária, financeira, gerencial e patrimonial, com o objetivo de fortalecer a capacidade de geração de receita própria e aprimorar os sistemas de informação, serviços e processos [13].

A avaliação ou escolha sobre quais clientes seriam objeto de novos negócios acontece nas IFs por meio de alguns critérios ou referenciais, ligados à relação risco X retorno, à expansão do *market share*, à diversificação de portfólio

e ao atendimento de requisitos regulatórios [16]. A combinação desses fatores e a percepção dos gestores constitui o processo decisório, consistindo em assumir ou não os riscos dos novos negócios e, por conseguinte, se beneficiar dos retornos. Todavia, em se tratando de entes públicos, há que se levar em conta o alinhamento entre o fomento ao crédito e o cumprimento das políticas públicas voltadas à promoção do desenvolvimento regional e social.

Nesse contexto, uma mensuração adequada do risco de crédito, combinada com a gestão dos negócios assumidos pelas IF junto aos entes públicos se faz essencial, para que os objetivos estratégicos da IF na gestão dos portfólios de crédito sejam atingidos. Assim surge o seguinte enunciado do problema de pesquisa: **Como mensurar o risco de crédito dos municípios brasileiros de forma massificada, sopesando aspectos diversos como nível de gestão pública, informações econômico-financeiras, características sociodemográficas e desenvolvimento social, de modo a utilizar a informação do risco no processo de gerenciamento dos riscos e negócios?**

1.3. Justificativa

Pesquisadores [17][18][19][20][21] destacam a necessidade crescente de avaliação do desempenho de governos locais, tornando-se uma grande questão em todo o mundo, dado que os municípios assumiram cada vez mais responsabilidades na prestação de serviços

essenciais aos contribuintes, nos últimos anos. Os resultados de estudos como esses podem ser do interesse de políticos, gestores, autoridades fiscais, governos centrais, órgãos de supervisão, instituições financeiras, bancos, eleitores, contribuintes e usuários de serviços públicos [19].

Outro ponto importante está no cumprimento dos requisitos regulatórios brasileiros e internacionais como as normativas relacionadas à Basileia e ao *International Financial Reporting Standards* nº 9 (IFRS 9). Essas normas trazem como requisito, a necessidade de que os ativos das IF tenham os seus riscos de crédito mensurados de forma probabilística, de modo a estimar uma perda esperada e inesperada relacionadas aos clientes e negócios. Tal requisito alcança todos os segmentos da IF, incluindo exposições junto a entes públicos como os municípios.

No contexto de uma IF, os *stakeholders*[1] esperam que exista uma busca constante voltada à sustentabilidade econômica e negocial. Para alcançar patamares de excelência, faz-se necessário investimento constante em desenvolvimento dos modelos para mensurar os riscos e gerenciar de portifólios de clientes e negócios, que vão além do cumprimento dos aspectos regulatórios [22].

O resultado possibilitará aos gestores de uma Instituição Financeira uma visão mais clara sobre o risco de crédito dos municípios, por meio de uma classificação de risco, associada a uma estimativa de probabilidade de *default*, que permita alinhar a estratégia de negócios e o apetite ao risco da IF. Dessa maneira, contribuindo para o processo de tomada de decisão e o atingimento dos objetivos estratégicos da IF.

1.4. Objetivos

1.4.1. Objetivo geral

O objetivo desta pesquisa é propor um modelo de decisão para classificar de forma massificada os municípios brasileiros, quanto ao risco de crédito, observando as características dos entes públicos, tais como: nível de gestão pública, informações econômico-financeiras, características sociodemográficas e desenvolvimento social.

[1] Exemplo de stakeholders em uma IF: funcionários, fornecedores, parceiros, clientes, credores, acionistas, concorrentes, comunidade, governo e entidades ligadas ao meio ambiente.

1.4.2. Objetivos específicos

Para alcançar o objetivo geral, destacam-se os seguintes objetivos específicos:

a) Caracterizar os aspectos necessários para avaliar e classificar os entes públicos

b) Desenvolver e aplicar um modelo de avaliação de risco de crédito no processo de ordenamento/classificação dos entes públicos;

c) Validar o modelo e os efeitos da sua utilização na classificação dos municípios.

1.5. Estrutura do trabalho

No capítulo 2, é realizada uma revisão sistemática da literatura, com o objetivo de melhor direcionar os esforços da pesquisa e garantir que trabalhos relevantes tenham sido considerados. Para isto, é utilizada a Teoria do Enfoque Meta-Analítico Consolidado (TEMAC). Nesse capítulo também é apresentada a fundamentação teórica dos assuntos correlatos ao objeto de pesquisa.

No capítulo 3, é apresentada a metodologia aplicada, com detalhamento do ambiente e recursos utilizados.

No capítulo 4, descrevemos o processo de modelagem e os modelos desenvolvidos.

No capítulo 5, realizamos validação dos modelos e apresentação dos resultados.

No capítulo 6, apresentamos as considerações finais, seguida das referências bibliográficas utilizadas.

2. Revisão de Literatura

2.1. Revisão do Estado da Arte

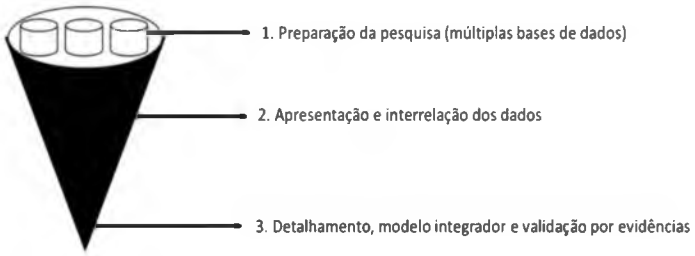
O grande volume de informação científica gerada a todo instante, tem o benefício de contribuir para a melhoria contínua de diversas áreas do conhecimento, dando suporte tanto ao processo de pesquisa e desenvolvimento, quanto à solução de problemas de diferentes naturezas. Por essa razão, quando se inicia um trabalho de pesquisa, é fundamental garantir a relevância do tema estudado, bem como considerar os trabalhos mais relevantes já desenvolvidos, para evitar aplicar esforços em estudos que não apresentem contribuição significativa para a ciência.

Neste ambiente de grande volume de dados, faz-se necessária a utilização de um método apropriado que auxilie o pesquisador na estruturação e análise dos trabalhos científicos disponíveis. Para atender a esta finalidade, a revisão do estado da arte foi realizada utilizando a Teoria do Enfoque Meta-Analítico Consolidado – TEMAC [23], que possui método estruturado e robusto, balanceando os esforços de sua aplicação face aos benefícios alcançados ao final do processo. Os itens a seguir descrevem as características do método, bem como os resultados específicos de sua aplicação para o contexto do risco de crédito dos municípios.

2.1.1. Descrição do método

Conforme Mariano e Rocha [23], a aplicação do TEMAC atende aos princípios do Enfoque Meta-Analítico, que utiliza os critérios de impacto de revistas, citações de autores e artigos e frequência de palavras-chaves, mas com o benefício de integrar ao processo ferramentas tecnológicas de acesso gratuito, que facilitam as análises e reduzem o trabalho manual.

Para atingir o resultado esperado, que é a identificação das referências mais relevantes para o contexto da pesquisa, tendo em vista o conhecimento prévio do que já foi publicado sobre o tema, o TEMAC conta com três etapas principais, conforme a Figura 2.1:



Fonte: Mariano e Rocha [22]

Figura 2.1: *Etapas do TEMAC*

a) Preparação da pesquisa: consiste na definição de palavras-chave relacionadas ao tema da pesquisa, o período de análise, as bases de dados utilizadas e as áreas de conhecimento que serão consideradas.

b) Apresentação e inter-relação dos dados: consiste em relacionar inúmeras fontes de informações, a critério do pesquisador, como a evolução do tema ano a ano, os autores mais citados, periódicos que mais publicam, entre outros.

c) Detalhamento, modelo integrador e validação por evidências: nesta etapa são identificados os principais autores, abordagens e linhas de pesquisa referentes ao tema, utilizando técnicas de co-citação e acoplamento (*coupling*).

Com base nos resultados dessas três etapas, o pesquisador deve então selecionar as referências que serão utilizadas para o desenvolvimento do trabalho. A metodologia TEMAC tem por objetivo destacar o que não deve faltar no trabalho, sem a pretensão de ser taxativo no sentido de limitar a abrangência das referências e fontes de pesquisa. Em última instância, a experiência do pesquisador, orientador e outros colaboradores deve ser levada em consideração.

2.1.2. Preparação da Pesquisa

2.1.2.1. Base de dados utilizada

A plataforma *Web of Science* (WoS) foi utilizada como referência para este estudo, em razão de sua reconhecida excelência operacional, da existência de plataforma própria de análise, que facilita a consolidação e extração dos dados, e da disponibilidade temporal a partir de 1945, garantindo maior cobertura do tema pesquisado. Os dados foram coletados considerando o período de 1945 a 2021.

Embora a plataforma contenha apenas publicações em inglês, esse fato não foi considerado uma limitação para esta pesquisa, já que o processo natural é que trabalhos

de elevada qualidade técnica sejam apresentados neste idioma, como forma de potencializar seu alcance a pesquisadores de diferentes partes do mundo.

2.1.2.2. Termos de pesquisa

Os termos de pesquisa utilizados buscam identificar publicações relacionadas a risco de crédito em municípios. Tendo em vista as diferentes possibilidades de caracterizar esses entes públicos, foi realizada uma pesquisa preliminar utilizando os seguintes termos: *Burgh, City, County, District, Judicial District, Metropolis, Municipal, Municipality, Town, Village e Local Government.* Associado a esses termos, foi inserido como critérios de busca o termo *Credit Risk.* Foi possível observar que as denominações de municípios que apareceram com maior frequência associada a risco de crédito foram *City, Cities, Muncipal e Municipality.*

Diante disso, adotou-se como parâmetros da pesquisa os termos e critérios detalhados no Quadro 2.1.

Quadro 2.1: *Combinação dos termos de pesquisa utilizados no TEMAC*

Objetivo de pesquisa	Termos Pesquisado	Campo Pesquisado
Município	*Cities* *City* *Municipal* *Municipality*	Título
Risco de Crédito	*Credit Rating* *Credit Score* *Default* *Financial Distress*	Todos

Fonte: Autoria própria

Para permitir a captura de variações dos termos pesquisados, foram utilizados asteriscos após o termo principal. O texto a seguir descreve a forma de busca utilizada a partir da ferramenta de pesquisa avançada da plataforma WoS:

$$((TI = (municipal\ or\ municipality\ or\ city\ or\ cities)\ AND\ TS = (default*\ or\ "credit*\ rating*"\ or\ "credit*\ scor*"\ or\ "financ*\ distress"))\ AND\ ALL = (credit))$$

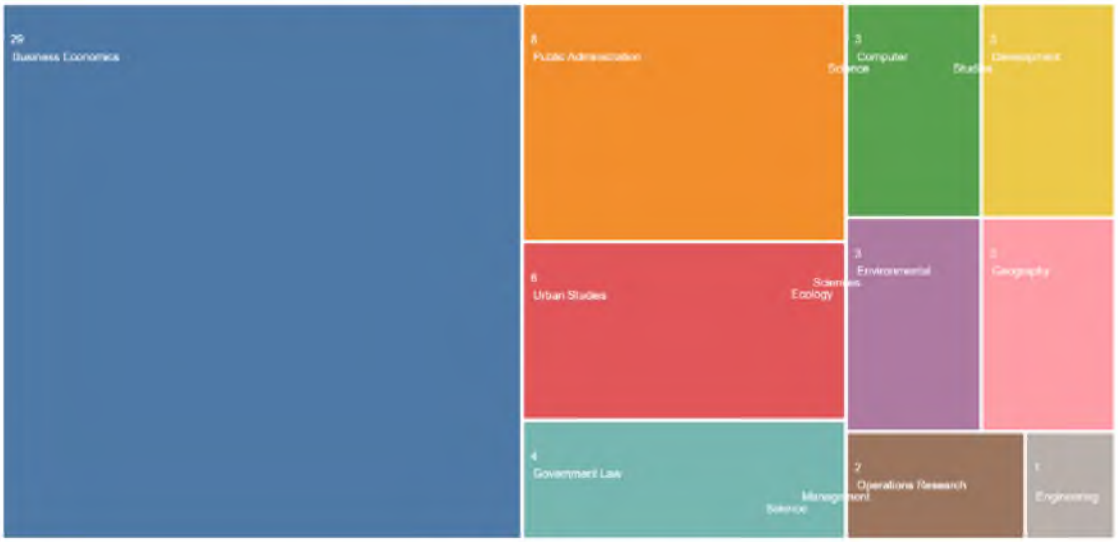
Onde TI refere-se à busca no título do documento, TS em todos os campos e ALL garante que todos os documentos contenham o termo crédito em alguma parte.

A pesquisa retornou 53 resultados, que serão detalhados nos tópicos seguintes.

2.1.2.3. Consolidação e tratamento dos dados para análise

De posse da base de dados com o resultado da pesquisa, foi possível obter as seguintes informações a respeito da natureza das publicações selecionadas.

a) Áreas de Pesquisa: a maior parte dos trabalhos são provenientes da área de pesquisa relacionada a Economia e Negócios, seguida das áreas de Administração Pública e Estudos Urbanos. A área de Computação está presente com apenas 3 publicações.



Fonte: Autoria própria

Figura 2.2: *Áreas de Pesquisa*

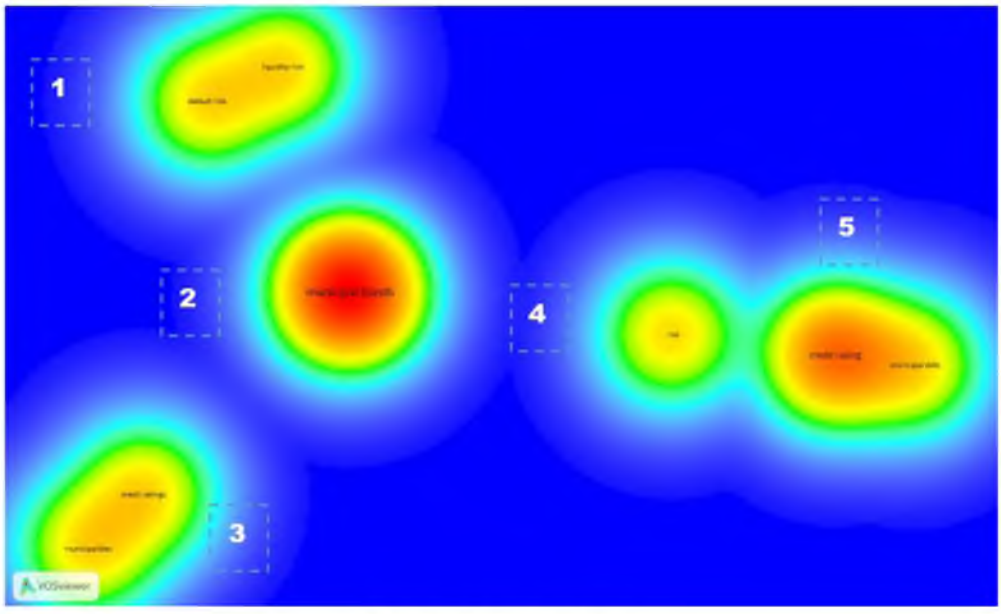
b) Áreas de conhecimento: embora a maior parte dos trabalhos estejam relacionados à área de Economia e Negócios, a avaliação das áreas de conhecimento revela que após a área de Ciências Sociais, a área de tecnologia é a mais presente nos estudos. Isso pode ser um indicativo de que ferramentas tecnológicas possam ser foco da abordagem dos trabalhos. A Figura 2.3 apresenta a relação das áreas de conhecimento.

Fonte: Autoria própria

Figura 2.3: *Áreas de Conhecimento*

c) Palavras-chave: a análise da frequência das palavras-chave revela os principais pontos de interesse presente nos trabalhos em cinco *clusters* principais, conforme apresenta a Figura 2.4. No cluster 1, percebe-se maior foco em risco de inadimplência e liquidez; no 2, os títulos públicos municipais aparecem em destaque. No cluster 3, destacam-se os termos rating de crédito e municípios. No cluster 4, o termo risco aparece de forma isolada, porém muito próximo do cluster 5, cujos termos rating de crédito e dívidas municipais aparecem juntos.

À primeira vista, a tendência é que os cluster 4 e 5 apresentem os trabalhos mais relacionados com o objetivo da pesquisa.

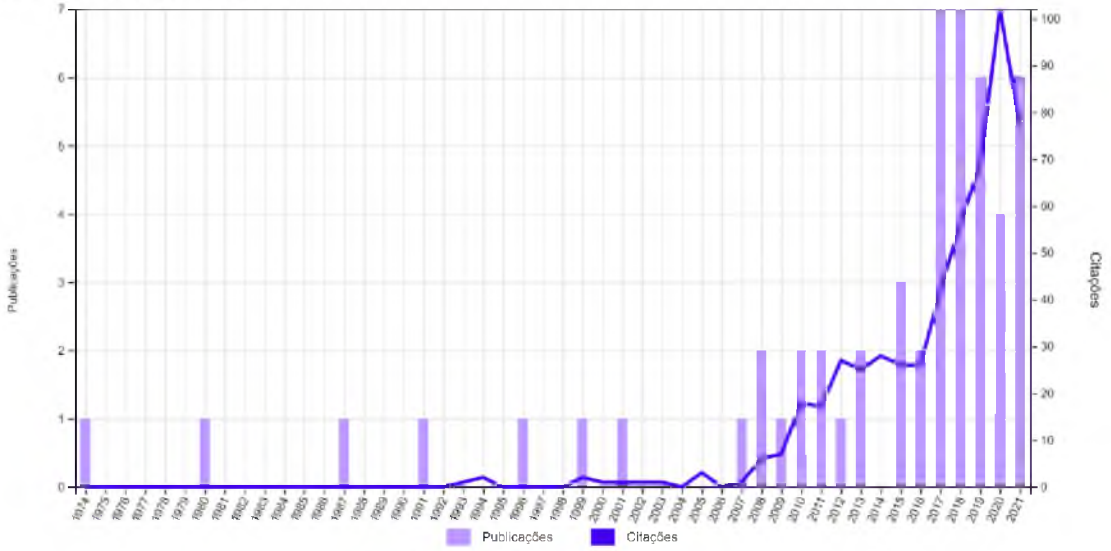


Fonte: Autoria própria

Figura 2.4: *Palavras-chave mais frequentes*

2.1.2.4. Autores e artigos mais citados

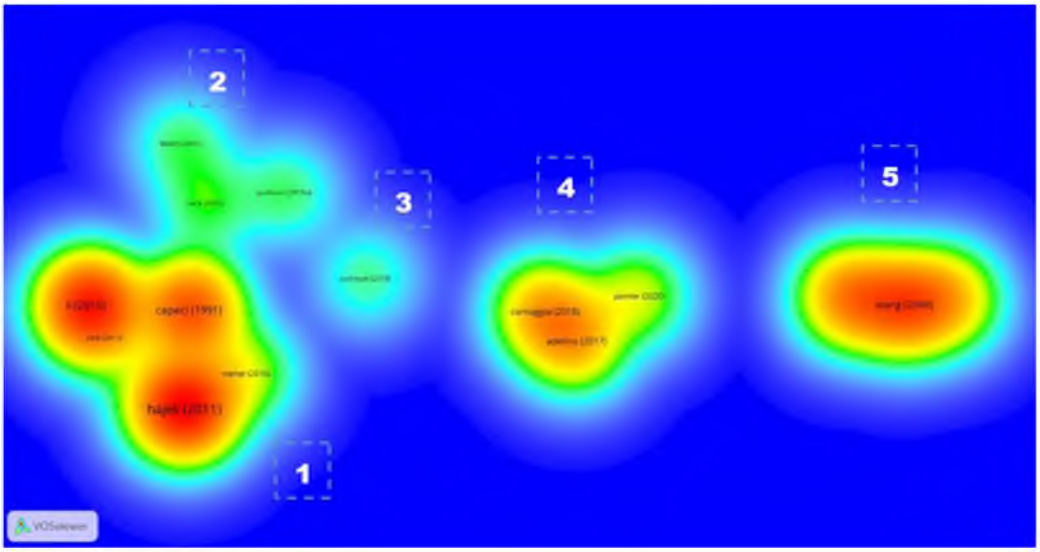
Ao longo dos anos, o número de publicações que contém os termos pesquisados se alterou de forma significativa. Nos anos iniciais, é possível observar intervalos de até sete anos entre publicações. Nos anos mais recentes o número de publicações tem se mantido em patamares de estabilidade, porém, com aumento no número de citações, conforme se observa na Figura 2.5.



Fonte: Autoria própria

Figura 2.5: *Número de Publicações e Citações ao longo do tempo*

Por meio da técnica de *coupling* ou acoplamento, é possível observar os documentos que são citados juntos com maior frequência dentre os trabalhos pesquisados. Essa avaliação revela as principais frontes de pesquisa identificadas na base de dados resultante.



Fonte: Autoria própria

Figura 2.6: *Coupling de documentos citados juntos*

Na Figura 2.6 são apresentados os 5 *clusters* identificados. Na sequência serão detalhados cada cluster com os autores e trabalhos em destaque.

2.1.2.5. Breve análise dos autores mais citados resultante da pesquisa TEMAC

No primeiro *cluster*, destacam-se os trabalhos de Capeci [24], Li [25], Hajek [26] e Alaminos [18].

Em 1991, Cappeci [24] utilizou a técnica de Dados em Painel para estimar a forma como a qualidade da gestão fiscal do município pode afetar suas taxas de empréstimos. O modelo econométrico evidenciou que tanto o risco de crédito quanto as taxas de juros ofertadas ao município têm relação direta com a qualidade dos indicadores fiscais do município, o que pode ser utilizado para impor alguma disciplina fiscal ao ente público.

Em 2010, Li [25] apresentou um trabalho com viés contábil, não relacionado aos objetivos deste estudo. Em seu trabalho, Li demonstrou que auditores com conhecimento a nível nacional agregam maior valor às empresas do que os profissionais a nível municipal.

Em 2011, Hajek [26] apresenta possibilidades de modelagem da classificação do risco de crédito de municípios americanos por meio de redes neurais. O trabalho apresenta de forma detalhada as etapas necessárias para a modelagem, incluindo o pré-processamento dos dados, técnicas para seleção de variáveis e a definição das estruturas da rede neural utilizada na classificação. O modelo é aplicado para municípios do estado de Connecticut, nos EUA. Um ponto importante é que o autor não modela a probabilidade de inadimplência dos municípios. Ele se utiliza da classificação de risco disponibilizada pela Moody's.

Em 2018, Alaminos [18] apresenta um estudo que utiliza mineração de dados, em que foram testados alguns métodos de *machine learning*, para previsão de problemas financeiros em municípios. Foram utilizados dados de 128 municípios espanhóis, sendo possível demonstrar a viabilidade do método para o objetivo proposto. Comparado com estudos anteriores, o modelo desenvolvido, por meio de redes neurais artificiais, neste trabalho aumenta a capacidade de prever dificuldades financeiras municipais e confirma que o uso de diferentes proxies da situação financeira de um município fornece resultados visivelmente diferentes.

No trabalho de Alaminos [18], entre as citações destacam-se as pesquisas de Cohen [17], Galariotis [21] e Gorina, 2017 [20]. A primeira, apresenta um modelo operacional para avaliar a viabilidade financeira de 360 municípios locais na Grécia, com base em uma metodologia multicritério, combinada com abordagem de análise de simulação (análise de aceitabilidade multicritério estocástica) com uma técnica de desagregação, utilizando como insumos índices financeiros personalizados para o contexto do governo local. O segundo, apresenta um modelo de avaliação financeira multi-atributo, que foi aplicado em toda a população dos municípios franceses no período de 2000–2012 para avaliar como as reformas implementadas na última década (tributação e descentralização) afetaram o desempenho financeiro dos governos locais na França. Já a terceira, em seu artigo, traz uma medida construída a partir de indicadores de dificuldade fiscal usando relatórios financeiros anuais abrangentes, orçamentos e cobertura da mídia. A técnica utilizada para estimar a *target* foi a Regressão Logística. A variável dependente escolhida foi em função do desempenho financeiro anterior e do ambiente socioeconômico. Os modelos empíricos mostram a importância relativa das reservas fiscais, dívida e composição da receita na previsão da dificuldade fiscal local.

No segundo cluster não há um trabalho dominante, porém os trabalhos citados juntos referem-se aos de Davies [27], Beck [28] e Padovani [29].

Em 2017, Davies [27] utilizou-se da técnica de Dados em Painel para modelar a influência de indicadores sociais e ambientais sobre os ativos líquidos municipais, tendo em vista que esses ativos influenciam o risco de crédito do município. Foi evidenciado que os ativos municipais são diretamente impactados pela incidência de crimes contra a propriedade, a riqueza da comunidade e a forma de governo da câmara municipal. Por outro lado, há uma relação indireta dos ativos com a passivo dos programas assistenciais, desemprego e crimes violentos. Os resultados sugerem que aspectos socioambientais não capturados por técnicas contábeis possam ser utilizados para melhorar a estimativa das variações nos ativos municipais.

Em 2018, Beck [28] aborda os riscos de que funcionários do governos atuem de forma imparcial na elaboração dos relatórios financeiros municipais, gerando impactos na interpretação utilização dos dados pela população, reguladores e pesquisadores. Os resultados, com teor contábil, demonstram que há assimetrias de informação que são utilizadas a depender do interesse municipal. Por exemplo, antes da emissão de títulos públicos, são utilizados artifícios contábeis para tornar os resultados favoráveis. O autor apresenta um modelo para comparação da saúde financeira entre cidades de diferentes países, novamente com alto teor contábil.

No mesmo ano, Padovani [29] alerta para o fato de que o sucesso do estabelecimento de uma nova empresa pode estar diretamente influenciado pela saúde

financeira do município no qual ela será sediada. Porém, esse aspecto é por vezes negligenciado por empresários, que se concentram em questões como menor carga de impostos, qualidade do sistema escolar e menor restrição regulatória.

De modo geral, é possível identificar que os trabalhos apresentados no segundo cluster possuem interesse por aspectos contábeis municipais.

O terceiro cluster refere-se ao trabalho de Padovani [30], que em 2018 estudou a relação entre o crescimento sustentável dos municípios italianos e seu endividamento mediante a emissão de títulos públicos. Foi demonstrado que os municípios que mais necessitam emitir títulos para promover suas políticas públicas são igualmente os que possuem piores indicadores sociais e, em contrapartida, acabam apresentando maior risco de crédito e taxas de juros piores. Seu trabalho busca identificar quais seriam as principais variáveis que contribuem para a melhoria da qualidade creditícia do município, sendo destacadas: melhoria da transparência na divulgação de informações financeiras, redução de atividade criminosa e maior frequência de auditoria externa.

O quarto cluster está centrado nos trabalhos de Adelino [31], Cornaggia [32] e Painter [33].

Em 2017, Adelino [31] demonstra o efeito sobre o emprego e o crescimento municipal quando o financiamento público é incentivado a partir da recalibragem do risco de crédito por meio de agências classificadoras de rating. Foi demonstrado que os municípios aumentam seu endividamento quando há elevação do rating, porque sua capacidade de assunção de dívidas se expande. Por outro lado, Adelino demonstrou que o aumento do investimento municipal gera um multiplicador estimado em 1,9 na renda local. A conclusão geral é que o aumento do financiamento da dívida pública pode melhorar as condições econômicas durante recessões, justamente no momento em que as instituições financeiras tendem a negar o crédito.

No ano seguinte, Cornaggia [32] se utiliza da mesma base de dados de Adelino, relacionada à reclassificação dos ratings municipais realizada pela Moody's em 2010. Ele questiona o fato do que esses ratings ainda são muito utilizados, embora não sejam atualizados com a qualidade que deveriam, tendo em vista o aumento de informações a respeito dos municípios e maior disponibilidade tecnológica. Sua sugestão é que houvesse menor dependência dessas classificações de ratings, que poderiam estar enviesadas por classificações ineficientes devido a padrões de classificação que variam entre diferentes classes de ativos.

Em 2020, Painter [33] estuda os efeitos das mudanças climáticas sobre o valor dos títulos públicos municipais. Sua constatação é que os municípios com maior propensão a sofrerem os efeitos das mudanças climáticas contam com taxas de juros mais altas para

financiamentos de longo prazo, quando comparados a municípios com menor risco de impacto climático.

O quinto e último cluster destaca o trabalho de Wang [34] e Schwert [35].

Em 2008, Wang [34] estuda a relação entre liquidez, inadimplência, impostos e rendimentos dos títulos municipais. Sua conclusão é que o risco de liquidez, muitas vezes ignorado quando da estimativa de risco de crédito, pode resultar em grave subestimação dos rendimentos dos títulos municipais.

Em 2017, Schwert [35] examina a precificação dos títulos públicos municipais, buscando entender a relação da liquidez do município e o valor dos ativos. A principal evidência apresentada no trabalho é que quando o governo local adota estratégias de isenção de impostos, o risco de crédito chega a responder por até 84% da redução do spread médio. O principal motivador seria o risco de diminuição da liquidez municipal.

De modo geral, pode-se concluir que a maior parte dos trabalhos avaliados possui viés contábil, ora focados na estrutura das dívidas públicas, ora na precificação de títulos municipais. Porém, os trabalhos de Alaminos [18], Gorina [20] e Hejek [26] apresentam abordagens que podem ser úteis para o desenvolvimento desse trabalho. Em seus estudos, são utilizadas técnicas computacionais relacionadas à Mineração de Dados com técnicas tradicionais estatísticas e de *machine learning*.

Percebe-se por meio dessa pesquisa, que apesar do grande avanço tecnológico e maior disponibilidade de informações nos últimos anos, ainda existe um gap científico em relação a abordagens que utilizam esse tipo de técnica para o objetivo de estimar a inadimplência municipal. A presente revisão do estado da arte contribuiu ainda, para reforçar a relevância do presente estudo em relação ao seu potencial de contribuir de forma efetiva com a ciência.

2.2. Referencial Teórico

O referencial teórico está subdividido em tópicos para esclarecer e orientar o desenvolvimento do estudo, à luz dos objetivos e resultados esperados. Nesse capítulo, serão tratados temas como gestão de riscos aspectos gerais e ligados à IF, assim como métodos computacionais relacionados à modelagem de dados fortemente desbalanceados (*"low default"*).

2.2.1. Princípios de Gestão de Riscos

Para efeito do estudo ora proposto, foi utilizado o referencial da norma ISO 31000:2009 – Gestão de Riscos – Princípios e Diretrizes [36], com suporte à norma IEC 31010:2009 – Gestão de Riscos – Técnicas de Avaliação de Riscos [37], as quais são

padrões globalmente aceitos para a gestão de riscos. Esses documentos são resultado de um trabalho desenvolvido por meio de um processo orientado e colaborativo, que envolveu centenas de profissionais de gestão de riscos ao redor o mundo ao longo de quatro anos.

Segundo Benetti [1], a criação da norma ISO 31000:2009 foi um dos efeitos positivos pós "Crise do Subprime". Após esse evento, a gestão de riscos emergiu como uma das atividades corporativas mais importantes, no sentido de ajudar a melhorar a saúde financeira das empresas. O padrão ISO 31000 oferece uma maneira simples de pensar sobre riscos e gerenciamento de riscos, organizando o modo de perceber os processos, orientando a resolução de problemas ponderando aspectos relacionados a riscos.

A ISO 31000 [38] destaca o propósito da gestão de riscos como um meio para criação e proteção de valor nas organizações, promovendo a melhoria contínua do desempenho, a busca constante da inovação e apoio para o alcance de objetivos organizacionais. Todavia, Lalonde e Boiral [39] destacam a importância sobre como a norma deve ser interpretada e implementada pelas organizações. Nesse sentido, a eficácia da ISO 31000 é, em última análise, determinada pela forma como ela é utilizada pelas organizações, e não apenas pela adoção ou não de sua estrutura de gerenciamento.

Para promover uma gestão de riscos eficaz e eficiente nas organizações, a ISO 31000 traz 8 (oito) princípios relacionados à criação e à proteção de valor, que devem nortear a estrutura e os processos de gerenciamento dos riscos nas entidades. A Figura 2.7 destaca quais são os princípios.

Fonte: ABNT [38]

Figura 2.7: *Princípios - ABNT NBR ISO 31000:2018*

2.2.1.1. Estrutura da Gestão de Riscos

Em sintonia com os princípios, a ISO 31000 recomenda que a estrutura de gerenciamento de riscos em uma organização seja integrada à governança e permeie todas atividades e funções finalísticas e de apoio. Nesse contexto, a alta administração nas organizações possui papel fundamental para que a gestão dos riscos seja eficaz, atuando em todas as dimensões da estrutura, desempenhando sua função de liderança de forma comprometida com a boa governança.

Destacam-se entre as dimensões observadas na estrutura:

a) Integração – a gestão de riscos integrada ao propósito organizacional, à governança, à liderança e comprometimento, à estratégia, aos objetivos e operações;

b) Concepção – a estrutura deve ser alinhada ao contexto externo e interno da organização, compreendendo propósito, responsabilidades, recursos, cultura, comunicação;

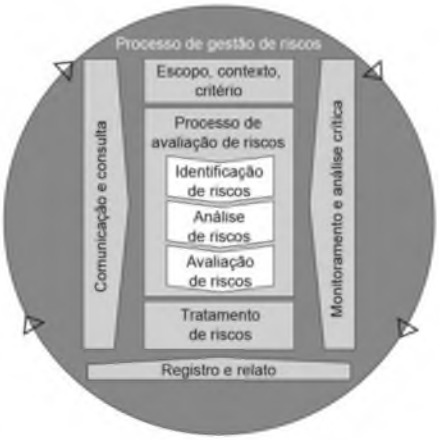
c) Implementação – convém que seja projeto estratégico para a organização, com planejamento articulado e envolvendo toda a organização;

d) Avaliação – a gestão de riscos avaliada periodicamente e sintonizada com o propósito e estratégia da organização;

e) Melhoria – gestão de riscos com melhoria contínua dos processos, com reflexo na estrutura e na integração.

2.2.1.2. Processo de Gestão de Riscos

Segundo a ISO 31000, a gestão de riscos nas organizações acontece por meio dos processos instituídos e suportados pela estrutura de gerenciamento dos riscos, com a aplicação sistemática de políticas, procedimentos e práticas para as atividades de comunicação e consulta, estabelecimento do contexto e avaliação, tratamento, monitoramento, análise crítica, registro e relato de riscos. A Figura 2.8, resume a proposta de visão do processo de gestão de riscos:



Fonte: ABNT [38]

Figura 2.8: *Processo de Gestão de Riscos - ABNT NBR ISO 31000:2018*

Um ponto chave da aplicação da Norma, para Purdy [40], é o processo de estabelecimento do contexto no qual a gestão de riscos será aplicada. Uma vez estabelecido de forma correta, todo o processo de avaliação dos riscos é feito de forma intuitiva e abrangente. Nesse sentido, os tópicos a seguir exploram em mais detalhes o contexto interno, externo e da gestão de riscos.

A Norma destaca, em certa medida, que em todas as atividades de uma organização há riscos envolvidos que devem ser gerenciados. Reforça ainda a importância de um processo de gestão de riscos como auxílio à tomada de decisão, ponderando incertezas, a possibilidade de cenários futuros, independentemente de serem intencionais ou não, e seus efeitos sobre os objetivos acordados.

Na dimensão do processo de avaliação de riscos, a ISO 31000 se desdobra em mais orientações e referências, consolidadas na ISO/IEC 31010 [37], voltadas à seleção e aplicação de técnicas sistemáticas para o processo de avaliação de riscos, dedicando maior espaço ao processo de gestão de riscos em uma organização, trazendo boas práticas na seleção e disponibilização de um conjunto de técnicas para o processo de avaliação de riscos, e tendo o cuidado de não trazer conceitos novos ou em evolução que não tenham atingido um nível satisfatório de consenso entre os especialistas.

2.2.1.3. Processo de Avaliação de Riscos

A parte central do processo de gestão de riscos está relacionada com a preparação e a condução da avaliação de riscos, chegando, conforme necessário, ao tratamento dos riscos. O processo começa por definir o que a organização quer alcançar e os fatores externos e internos que podem influenciar o sucesso em alcançar esses objetivos. Este

passo é chamado de "estabelecer o contexto" e é um insumo essencial para a identificação de riscos [40].

O processo de avaliação dos riscos da Norma está estruturado em três fases:

a) Identificação de riscos – atividade está relacionada a encontrar, reconhecer e registrar os riscos na organização ou processos;

b) Análise de riscos – o objetivo dessa atividade é entender o risco, os quais necessitam ser tratados, e identificar estratégias e métodos para tratamento mais adequados à situação;

c) Avaliação de riscos – é nessa atividade que a organização, a partir do contexto definido, avalia os níveis estimados de risco e estabelece tratamento mais adequado ao apetite aos riscos expostos.

Segundo a ISO/IEC 31010, o processo de avaliação de riscos na organização pode ser conduzido em vários níveis de profundidade e detalhe, utilizando um ou muitos métodos, que variam do simples ao complexo, conforme o contexto ao qual as técnicas serão aplicadas.

2.2.1.4. Processos de Mensuração de Riscos

Entre os modelos utilizados para orientar um processo de mineração de dados, destaca-se o uso do modelo *CRoss Industry Standard Process for Data Mining* (CRISP-DM) [41]. Em recente trabalho de revisão sistemática da literatura, publicado em 2021 [42], chegou-se à conclusão que o CRISP-DM é considerado o padrão de fato em projetos de mineração de dados, sendo ainda um modelo de processo adequado, mesmo após 20 anos do seu lançamento.

O CRISP-DM consiste em um conjunto de processos estruturados, com o objetivo de orientar a mineração de dados para resolução de problemas. A seguir são destacados os processos do CRISP-DM.

a) Entendimento do Negócio: conhecer o ambiente e o que a organização espera da mineração de dados. Faz-se necessário envolver as pessoas-chave no entendimento do negócio e nas discussões. Nesta fase é construído um plano de projeto para atingir os objetivos;

b) Entendimento dos Dados: a coleta e conhecimento inicial dos dados acontecesse nessa fase é realizado um olhar mais atento aos dados disponíveis para mineração, que podem ser dados existentes da organização ou dados comprados;

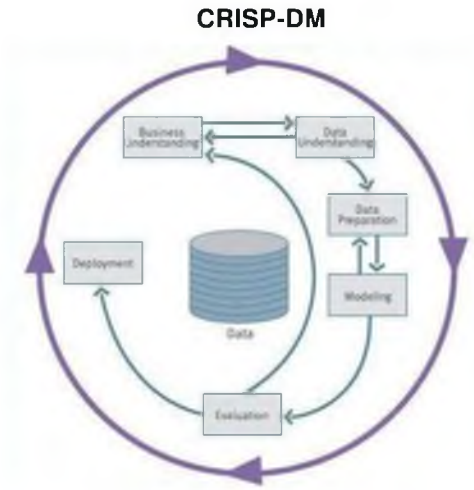
c) Preparação dos Dados: é um dos aspectos mais importantes e normalmente demorado na mineração de dados, ao final é construída a base de dados final que será utilizada na fase seguinte;

d) Modelagem: nesta fase são executados vários classificadores e algoritmos, gerando diversos modelos, utilizando os parâmetros padrão ou ajustados dos algoritmos;

e) Avaliação: após o desenvolvimento dos modelos, faz-se necessário avaliar os resultados usando os critérios de sucesso dos negócios estabelecidos no início do projeto;

f) Implantação: nesta fase há efetivamente o uso dos modelos nos processos, que pode ser colocar em produção, por exemplo em alguma aplicação de sistema ou ferramenta.

Entre as fases do CRISP-DM, destaca-se a do "entendimento do negócio", sendo esta relacionada diretamente ao objetivo do estudo. A Figura 2.9 mostra o processo e as fases do CRISP-DM numa perspectiva de fluxo.



Fonte: IBM [41]

Figura 2.9: *Modelo CRISP-DM*

2.2.2. Gestão de Riscos em Instituições Financeiras

Historicamente, as iniciativas formais a respeito do assunto "regulação bancária" começaram a existir no início dos anos 70 e resultaram na constituição do Comitê de

Supervisão Bancária de Basileia, um fórum vinculado ao BIS, formado por representantes dos bancos centrais e autoridades com responsabilidade formal sobre supervisão bancária do chamado G10[2].

O Comitê de Basileia, possui como características: não fazer parte do BIS, não possuir autoridade formal de supervisão internacional e suas conclusões não ter caráter de força legal. Seu objetivo é a elaboração de padrões de supervisão, bem como recomendações e princípios para as melhores práticas no mercado financeiro, na expectativa de que as autoridades de cada país adotem medidas para implementá-las, e que sejam ajustadas às suas realidades locais.

Com o estabelecimento de padrões para possibilitar o aprimoramento da qualidade da supervisão bancária, o Comitê de Basileia contribuiu para reduzir as desigualdades competitivas entre instituições financeiras internacionalmente ativas e, ainda, fortalecer a solidez e a segurança do sistema bancário internacional.

O primeiro Acordo de Capital de Basileia, aprovado em 1988[43], recomenda padrões mínimos de requerimento de capital para fazer frente aos riscos que deterioram os capitais dos bancos internacionais na década de 80. O foco principal deste acordo foi a ponderação dos ativos sujeitos ao risco de crédito.

Os requerimentos mínimos de capital são estabelecidos pelos reguladores para que as IF tenham níveis mínimos de capital para fazer frente às perdas inesperadas, no sentido de se buscar a estabilidade do sistema financeiro, com redução do risco de a sociedade arcar com custos advindos de instituições insolventes.

O Acordo de Capital de 1988[43] foi pioneiro em estabelecer requerimentos de adequação de capital baseados em risco de crédito. Suas recomendações estipulavam um capital mínimo regulatório, contemplando a exigência de capital de 8% sobre os ativos ponderados a um fator de risco, que variava em função da liquidez e do risco associado. Esses ativos basicamente são empréstimos ou outros ativos monetários, prestação de garantias (ativos *off-balance*) e contratos de derivativos.

A sua implementação foi bombardeada de críticas do mercado e não faltaram novas recomendações com o fito de refletir, entre outras mudanças, a crescente concorrência dentro dos mercados financeiros, as práticas sofisticadas de gestão de riscos

[2] Alemanha, Bélgica, Canadá, Espanha, Estados Unidos, França, Inglaterra, Holanda, Itália, Japão, Luxemburgo, Suécia e Suíça.

e as frequentes inovações financeiras. Das críticas feitas ao Acordo, destacam-se quatro problemas principais:

a) assimetria no tratamento do risco de crédito: observou-se um favorecimento às instituições de países membros da Organização para Cooperação e Desenvolvimento Econômico (OECD), com a aplicação de pesos diferenciados entre os países membros, considerados menos arriscados, e os não-membros, considerados mais arriscados;

b) não havia qualquer vantagem, como redução da carga de capital exigido, como incentivo às melhorias na gestão das instituições financeiras;

c) o risco considerava as peculiaridades das carteiras de ativos e não observava a diversificação ou a concentração das carteiras e não havia provisão para posições vendidas;

d) o Acordo de 1988 previa apenas valores de exigência de capital para risco de crédito, deixando de fora outros riscos.

Uma das recomendações se materializou em janeiro de 1996, com a publicação de documento intitulado *Amendment to Capital Accord to Incorporate Market Risks*[44], em que foi adicionada, além de requerimentos de risco de crédito, a exigência de capital para suportar riscos de mercado. Na verdade, tratava-se de um adendo ao Acordo de 1988 que, apesar de recomendar a utilização de uma abordagem padrão, facultava aos bancos a utilização de modelos internos, desde que validados pelo órgão regulador local. Portanto, o índice mínimo de capital para o ativo ponderado pelo risco, passou a incorporar o risco de mercado. Particularmente, no âmbito do Brasil nem todas as recomendações relacionadas ao risco de mercado foram implementadas, por exemplo: risco de ações, risco de taxa de juros variável, risco de commodities e outros.

2.2.2.1. Acordo de Capitais – Basileia II

Apesar das diversas atualizações (emendas) a ineficácia latente do Acordo de Basileia I, principalmente em decorrência de práticas de arbitragem de capital e da crescente sofisticação das ferramentas de gestão de riscos, provocou o movimento de revisão estrutural na forma de apurar, controlar e gerir os riscos das instituições financeiras. O Comitê de Basileia, após vários estudos em parceria com o mercado financeiro e acadêmico, publicou o Novo Acordo de Capital, titulado Basileia II, em junho de 2004 [5].

No Basileia II, o objetivo, destacado pelo Comitê de Basileia, foi promover o desenvolvimento do sistema financeiro internacional como um todo, no sentido de fortalecer a solidez e a estabilidade do sistema, por meio de uma gestão de riscos mais forte, além de manter consistência suficiente para que a regulamentação da adequação de capital não fosse uma fonte significativa de desigualdade competitiva entre as instituições internacionalmente ativas.

A nova forma de enxergar proposta pelo Comitê de Basileia representa um avanço significativo em relação ao modelo vigente até então, na medida em que faculta às instituições financeiras desenvolver modelos proprietários para cálculo da exigência de capital para risco de mercado e operacional e modelos internos dos componentes de risco de crédito, considerados insumos do motor de cálculo de exigência de capital para esse risco. Observa-se que para risco de crédito o modelo de apuração da exigência de capital é dado pelo Basileia II.

O Novo Acordo[5] se propõe a aproximar as dimensões de capital regulatório[3] e capital gerencial[1], tornando o capital regulatório mais sensível aos níveis de risco presentes nas carteiras de crédito dos bancos. Nesse contexto, essa inovação de Basileia II observa uma das recomendações verificadas pelo mercado, para melhoria da estrutura regulatória.

Em linhas gerais, a estrutura do Novo Acordo está calcada sobre três pilares:

a) Pilar I – Exigência Mínima de Capital
b) Pilar II – Supervisão Bancária
c) Pilar III – Disciplina de Mercado

A estrutura apresentada em Basileia II provê metodologias e parâmetros para que os bancos identifiquem, quantifiquem e controlem seus riscos ao invés de se impor um único modelo para o cálculo dos requisitos de capital, que possibilita tornar a exigência de capital regulatório mais sensível aos riscos subjacentes aos negócios.

Uma das inovações trazidas por Basileia II, refere-se ao estabelecimento de requerimentos mínimos de capital para fazer frente aos riscos operacionais, representando um grande desafio à sua mensuração. Na verdade, a inclusão de exigência de capital para risco operacional objetiva incentivar melhorias na gestão das instituições financeiras.

---

[3] Valor de referência exigido como piso de capital pelo regulador e apurado com base nos parâmetros regulatórios.

[1] Valor de referência apurado pela instituição financeira que serve de piso para o capital próprio da instituição financeira, com base em modelos e metodologias proprietárias.

Pilar I – Exigência Mínima de Capital

O Basileia II aprimora a forma de apurar o nível de capital mínimo das instituições financeiras, levando em consideração a complexidade das atividades das instituições financeiras. A forma como o Acordo foi proposto, leva as instituições financeiras a buscar continuamente melhorias nos processos de gestão de riscos, com a aplicação de metodologias cada vez mais sensíveis ao risco, que gerem, por conseguinte, requisitos de capital mais acurados e aderentes às características das carteiras de ativos.

Entre as inovações do Basileia II, no contexto do Pilar I, destacam-se a exigência para risco operacional e o reconhecimento de metodologias e modelos mais sofisticados para mensuração do risco de crédito. Os requerimentos para medir o risco de mercado e a exigência mínima de 8% de capital para os ativos ponderados pelo risco permanecem inalterados.

Risco de Crédito

Para mensuração da exigência de capital mínimo para fazer frente ao risco de crédito são apresentadas duas metodologias de apuração:

a) Abordagem Padronizada – a modelo é semelhante ao do Acordo anterior, no entanto apresenta-se mais sensível ao risco de crédito. A instituição poderá se valer de uma agência pública ou privada de classificação de risco, como a *Moody's, Standard & Poor's, Fitch* ou outras agências de *rating* especializadas em medir o risco dos ativos. O Basileia II estabelece um peso de risco para cada tipo de contraparte ou titular do crédito a ser ponderado na exposição (empréstimos e financiamentos), diferentemente do Acordo em vigor que lineariza a forma de tratamento, aplicando apenas uma única categoria de fator de risco a ser aplicado na exposição: 100%. Segundo a abordagem padronizada, os ativos dos bancos são classificados nas faixas de risco de acordo com a classificação de risco de crédito do devedor, elaborado pelas agências de *rating*. Para obter o requerimento mínimo de capital para risco de crédito, toda a exposição de crédito, conhecida como *Exposure at Default* (EAD), em cada faixa de peso de risco são somadas e multiplicadas pelo peso de risco apropriado e então multiplicadas pelo requerimento de capital total de 8%.

b) Classificação na Abordagem IRB – Permite às IF a se valerem de estimativas internas para os três componentes do risco de crédito, a saber: (i) PD – *Probability of Default* –, refere-se a probabilidade de um tomador de crédito ou contraparte incorrer em inadimplemento em um determinado período de

tempo, por exemplo um ano; (ii) LGD – *Loss Given Default* – está associada a perda efetiva apurada após processo de recuperação do crédito, como por exemplo o valor em prejuízo apurado pós liquidação parcial de um empréstimo inadimplente, por meio da incorporação pela instituição financeira de uma garantia vinculada ao crédito; (iii) EAD está associada ao montante de créditos (exposições) de um tomador ou contraparte no momento em que se verifica a inadimplência. Esses componentes de risco são utilizados como insumos do motor de cálculo do capital mínimo necessário, para fazer frente ao risco de crédito. Existe também um parâmetro colocado por Basileia II, chamado de *Maturity* (M), que deve ser calculado e informado quando da apuração do capital.

Existem duas variantes de IRB disponíveis para os bancos:

a) Básica: somente a PD proprietária, ou seja desenvolvida pelas IF. A LGD é fixada e baseada em valores do supervisor; Basileia II sugere 45% ou 75% (Basileia 2004, parágrafos 287 e 288). A EAD é também baseada em valores do Regulador, nos casos em que a mensuração não é clara, e a M é em média 2,5 anos (Basileia 2004, parágrafos 318, 319 e 324).

b) Avançada: os três componentes mais o parâmetro M são desenvolvidos e estimados pela instituição financeira, por meio de modelos internos relacionados aos clientes e exposições, abrangendo vários públicos e segmentos na IF, como por exemplo: pessoas físicas, pessoas jurídicas, produtores rurais, bancos, entes públicos etc.

Para as abordagens IRB, as instituições financeiras devem seguir normas mais rígidas de avaliação e fornecer maior transparência ao mercado. Seu uso, porém, depende de aprovação prévia da autoridade reguladora do país, no caso do Brasil o Banco Central do Brasil. Basileia II considera também outros aspectos no ambiente das instituições financeiras como por exemplo: modelos de classificação de clientes (ex.: *credit scoring, behaviour scoring, collection scoring*), tratamento de garantias reais e fidejussórias, aplicação de *haircuts* nos mitigadores financeiros, acordos de compensação (*netting agreement*), securitização de ativos.

Pilar II – Supervisão Bancária

A dimensão de supervisão bancária (Pilar II) está estruturada em premissas relacionadas ao supervisor, no que se refere à regulamentação, supervisão e acompanhamento dos riscos sistêmicos e também a questões vinculadas às instituições financeiras como governança, gestão de negócios e riscos. No Pilar II, observa-se como as

IF estão preparadas para identificar, medir e gerir tanto os riscos descritor no Pilar I, quanto outros que podem impactar negativamente ou positivamente seus negócios e, por conseguinte, sua estrutura de capital.

Para as instituições financeiras atenderem os requisitos do Pilar II, faz-se necessário o desenvolvimento e implementação de uma estrutura de gestão de riscos compatível com os riscos a que estão expostas e com a complexidade de suas atividades. Ao supervisor cabe avaliar nas instituições financeiras a qualidade dos processos de gestão de riscos e, quando necessário e de modo tempestivo, intervir com o fim de determinar melhorias e promover revisões técnicas de alocação de capital, objetivando a estabilidade e segurança no sistema financeiro.

O Acordo de Basileia II estabeleceu quatro princípios, no Pilar II, citados a seguir:

a) os bancos disponham de processos para avaliar seu capital global em relação ao perfil de risco de suas posições;

b) as autoridades de fiscalização devem examinar as avaliações e estratégias da adequação do capital das instituições;

c) a fiscalização espera que os bancos operem acima dos índices mínimos de capital regulador;

d) as autoridades de fiscalização devem procurar intervir em estágio inicial.

Esses princípios estão relacionados à existência de processos para identificação, mensuração e monitoramento de riscos e de determinação das necessidades de capital nas instituições financeiras. Já sob a perspectiva da supervisão, esses processos deverão ser avaliados sob aspectos qualitativos e quantitativos, bem como a adequabilidade dos níveis de capital.

No Pilar II, é estabelecida uma relação direta (embora não automática) entre o capital econômico e o regulatório [45]. Entre os quatro princípios citados, notadamente, dois se referem diretamente ao capital econômico. O primeiro princípio estabelece que as instituições financeiras devem colocar em prática um processo de determinação e manutenção de nível de capital compatível com seu perfil de risco, considerando todos os riscos assumidos pela IF (risco de crédito, risco de mercado, operacional risco, risco de taxa de juro da carteira bancária, liquidez risco, e outros riscos, tais como reputação e estratégico). O outro princípio está relacionado ao estabelecimento de estratégia para manter o nível adequado de capital econômico acima do mínimo, no caso a referência regulatória.

Pilar III – Disciplina de Mercado

O Pilar III foi uma novidade trazida por essa versão do Acordo e tem como objetivo promover a disciplina de mercado, por meio da divulgação de informações importantes das instituições financeiras para o mercado. O Pilar III estabelece critérios mínimos de transparência qualitativos e quantitativos suficientes para que o mercado possa formar opinião sobre como a instituição financeira gere seus riscos.

A intenção do Basileia II foi diminuir a assimetria de informação entre as instituições financeiras e o mercado, nesse sentido o que se espera é quanto maior o nível de informações e dados divulgados para o mercado, retratando o modo como as instituições financeiras atentam para os riscos no ciclo de gestão (identificar, medir e gerir), menor a possibilidade de os investidores, depositantes, sociedade, absorverem perdas decorrentes de situações de insolvência.

O Banco Central do Brasil (Bacen), órgão regulador do Sistema Financeiro Nacional (SFN), recepcionou as recomendações internacionais, com as devidas adaptações ao Brasil e publicou uma série de normativos: resoluções, circulares e carta-circulares, abordando a gestão dos vários riscos que afetam às IF. Os padrões relacionados aos modelos internos para mensuração do risco de crédito foram internalizados pelo Bacen, por meio da Circular nº 3.648, de 04 de março de 2013 [46].

2.2.2.2. Pós – Basileia II

A indústria financeira historicamente, por conta do alto nível de regulamentação e da necessidade de gerenciamento dos seus riscos, faz uso de ferramentas estatísticas, algoritmos matemáticos e soluções tecnológicas, chamados de modelos, para mensurar impactos dos riscos na sua atividade [47][48]. Suas aplicações são as mais variadas, internamente nas IF essas soluções são utilizadas nas áreas especializadas em clientes, finanças, risco, capital, produtos e serviços, auxiliando, por exemplo, nas tomadas de decisão em atividades de prospecção de novos clientes, de mensuração de perdas esperadas de crédito, de mensuração do retorno ajustado ao risco.

O desenvolvimento de modelos tornou-se um processo estratégico para as IF, cujo instrumento é considerado de extrema importância para o processo decisório, dada relevância e impacto nos negócios. Portanto, a forma como o desenvolvimento dos modelos é estruturada e executada reflete diretamente na qualidade desses instrumentos, que podem interferir, dado o nível de acurácia e outros indicadores de desempenho, nas decisões em todos os níveis da organização, [47] gerando mais ou menos riscos e consequências danosas para as instituições.

A estratégia de gestão de portfólio nos bancos passou por transformações por conta da crise financeira do "Subprime" de 2008. Segundo Hamerle et al.[16], a incorporação de várias características relacionadas a fatores macroeconômicos passou a ter uma importância maior nos modelos aplicados para gerenciamento da carteira de crédito, no sentido de melhor planejar as ações dados os cenários futuros, inclusive em ambientes estressados.

Uma iniciativa, também decorrente da Crise de 2008, foi protagonizada pelo *International Accounting Standards Board* (IASB), organismo internacional responsável pela publicação e atualização de normas contábeis, que publicou, em 2014, o *International Financial Reporting Standards* nº 9 (IFRS 9), novo padrão contábil para reconhecimento das perdas decorrentes do risco de crédito [49]. A proposta do IASB trouxe uma nova forma de mensuração das Perdas Esperadas de Crédito (ECL), mais alinhada à perda esperada regulatória do Comitê de Basileia, proposta nos modelos internos de Basileia II [50]. Esse padrão traz como novidade o reconhecimento antecipado das perdas de crédito, reduzindo o acúmulo de excesso de provisões e a superestimação da necessidade de capital regulatório, quando o crédito está mais deteriorado [51].

O padrão IFRS 9 não está regulamentado na indústria bancária brasileira, o Bacen colocou em discussão o tema por meio do Edital de Consulta Pública nº 60/2018, não havendo até o momento normas publicadas. Todavia, a observação quanto aos requisitos já é obrigatória para as IF que possuem operações fora do Brasil, por estar vigente desde janeiro de 2018. O IFRS 9 altera a relação entre os ativos considerados problemáticos (NPL) e as provisões para fazer frente às perdas esperadas (ECL). No processo de implementação nos países, observa-se variações no tratamento das inadimplências, nas adaptações dos regimes contábeis e na aplicação pelas próprias empresas em suas escriturações, gerando diferenças de conceitos, com reflexo no padrão contábil e no que é considerado efetivamente como um Ativo Problemático [52] no portfólio das IF.

Em 2015, a Associação Internacional de Gestores de Carteira de Crédito e *Price-Waterhouse Coopers* (PwC) realizou em conjunto o estudo mais importante até o momento para entender as práticas e os desafios da indústria financeira no desenvolvimento, implementação e aprimoramento do *Risk Appetite Framework* (RAF). A RAF é uma abordagem geral para estabelecer, comunicar e monitorar todos os riscos materiais de uma instituição financeira, por meio de papéis e responsabilidades organizacionais, declarações de apetite aos riscos, políticas, limites de riscos, processos, controles e sistemas. Esse framework orienta entre outras ações, as relacionadas à gestão de portfólios, subsidiando os processos decisórios de assunção de riscos no gerenciamento dos negócios [53].

Ainda no contexto dos reflexos da crise financeira de 2008, na indústria financeira, vários bancos procuraram fortalecer as suas estruturas de apetite ao risco (RAF). Como resultado da referida pesquisa [53], há o consenso entre os Supervisores e as entidades supervisionadas de que a RAF é essencial no processo de governança de riscos e negócios e deve estar em contínuo aperfeiçoamento. Segundo os pesquisados, o estabelecimento de uma RAF mais bem estruturada levou a uma conscientização de risco mais abrangente, chegando em todos os níveis organizacionais, e a uma melhor compreensão interna do perfil de risco da organização, e assim possibilitando um melhor alinhamento entre o apetite de risco e as metas estratégicas da organização. Esse entendimento é corroborado pela *International Association of Credit Portfolio Managers* (IACPM) em pesquisa realizada, em 2019 referente ao ano base de 2018, junto a 60 empresas líderes globais, para analisar a evolução do gerenciamento dos portfólios sujeitos ao risco de crédito, estruturas organizacionais, missão e mandato, ferramentas e perspectivas para o futuro [22].

2.2.3. Métodos computacionais aplicados na modelagem de bases desbalanceadas

2.2.3.1. Técnicas de tratamento de dados desbalanceados

O tratamento de dados desbalanceados é um tema muito estudado e há vários pesquisadores ao redor do mundo preocupados em entender e tratar o forte desbalanceamento de classes em mineração de dados [54][55][56][57][58][59][60][61][62]. Esse problema é encontrado em diversas frentes de aplicação no mundo real como detecção de fraude, detecção de *spam*, previsão de rotatividade, predição de *default*, detecção de anomalias, detecção de *outliers* entre outras. Entre as técnicas computacionais que se destacam como caminhos possíveis para minimizar os efeitos de desbalanceamentos acentuados são [59]:

a) Métodos de *oversampling*: são técnicas utilizadas para "duplicar" observações ou "sintetizar" novos registros a partir dos exemplos da classe minoritária. Seguem alguns dos métodos mais amplamente utilizados em mineração de dados:

   i. *Random Oversampling*: método mais simples que gera duplicação aleatória de exemplos da classe minoritária, no conjunto de dados de treinamento;

   ii. *Synthetic Minority Oversampling Technique* (SMOTE): considerado o método mais popular, seu funcionamento se dá pela seleção de exemplos

que estão próximos no espaço da variável minoritária, desenhando uma linha entre os exemplos no espaço do recurso e desenhando uma nova amostra como um ponto ao longo dessa linha.

iii. Borderline-SMOTE: seleciona as instâncias da classe minoritária que são classificadas incorretamente, como um *k-nearest* modelo de classificação de vizinhos, um único gerador de amostras sintéticas difíceis de classificar.

iv. *Borderline Oversampling with* SVM: é uma extensão para SMOTE que ajusta um *Suport Vetor Machine* (SVM), para o conjunto de dados. Utiliza o limite de decisão definido pelos vetores de suporte, como base para gerar dados sintéticos, com base na ideia de que o limite de decisão é a área onde mais exemplos minoritários são necessários.

v. *Adaptive Synthetic Sampling* (ADASYN): é outra extensão do SMOTE que gera amostras sintéticas inversamente proporcionais à densidade dos exemplos na classe minoritária. Essa técnica foi projetada para criar observações sintéticas em regiões do espaço de recursos onde a densidade de exemplos minoritários é baixa, e menos ou nenhum onde a densidade é alta.

b) Métodos de *undersampling*: são técnicas que excluem ou selecionam um subconjunto de observações da classe majoritária. A seguir estão listados alguns dos métodos mais usados:

i. *Random Undersampling*: método simples que promove a exclusão aleatória de observações da classe majoritária no conjunto de dados de treinamento.

ii. *Condensed Nearest Neighbor Rule* (CNN): método projetado para reduzir a memória necessária para o algoritmo de vizinhos k-mais próximos. Ele funciona enumerando as observações no conjunto de dados e adicionando-as à amostra apenas se elas não puderem ser classificadas corretamente pelo conteúdo atual da amostra. A técnica pode ser aplicada para reduzir o número de observações da classe majoritária, após todas as observações das classes minoritárias serem adicionadas na amostra.

iii. *Near Miss Undersampling*: uma família de métodos que usam *K-nearest neghbors* (KNN) para selecionar exemplos da classe majoritária.

iv. *Tomek Links Undersampling*: talvez o método mais conhecido, originalmente desenvolvido como parte da extensão CNN. A abordagem se referência em um par de observações no conjunto de dados de treinamento que são vizinhos mais próximos (têm a distância mínima no espaço da variável) e pertencem a classes diferentes, para promover a exclusão de observações da classe majoritária.

v. *Edited Nearest Neighbors Rule* (ENN): outro método para selecionar exemplos para exclusão. Essa técnica envolve o uso de k = 3 vizinhos mais próximos para localizar as observações em um conjunto de dados que estão classificados "incorretamente", excluindo-os da amostra.

vi. *One-Sided Selection* (OSS): é uma técnica que combina o *Tomek Links* e CNN. Nessa abordagem, o *Tomek Links* é usado para remover exemplos "barulhentos" no limite da classe, enquanto o CNN é utilizado para remover exemplos redundantes do interior da densidade da classe majoritária.

vii. *Neighborhood Cleaning Rule* (NCR): é outra técnica que combina CNN, para remover exemplos redundantes, e ENN para remover exemplos ruidosos ou ambíguos.

c) Métodos combinados de *oversampling* e *undersampling*: Embora métodos de *oversampling* ou *undersampling*, quando aplicados individualmente em um conjunto de dados de treinamento, possam já apresentar bons resultados, experimentos realizados [61][62] mostraram que a aplicação conjunta desses métodos pode resultar em melhores desempenhos de modelos ajustados à amostra de dados transformada. Algumas das combinações mais amplamente utilizadas:

i. SMOTE and *Random Undersampling*

ii. SMOTE and *Tomek Links*

iii. SMOTE and *Edited Nearest Neighbors Rule*

São métodos combinados em que são removidas observações com "ruídos" ao longo da fronteira de ambas as classes. No caso dos itens "ii" e "iii", primeiramente é aplicado o método SMOTE, na sequência são excluídas as observações pelos métodos *Tomek Links* ou KNN, respectivamente.

2.2.4. Abordagem sensível ao custo

O aprendizado sensível ao custo é uma subárea de pesquisa ligada ao aprendizado de máquina, em que os custos de erros de previsão (e potencialmente outros custos) são considerados no treinamento de algoritmos. Como explica Brownlee [59], há uma relação estreita entre a classificação desequilibrada e o aprendizado sensível ao custo, assim um problema de aprendizagem desequilibrada pode ser resolvido aplicando a aprendizagem sensível ao custo. No entanto, a aprendizagem sensível ao custo é um subcampo de estudo separado e o custo pode ser definido de forma mais ampla do que o erro de previsão ou erro de classificação. Isso significa que, embora alguns métodos de aprendizagem sensível ao custo possam ser úteis na classificação desequilibrada, nem todas as técnicas de aprendizagem sensível ao custo são técnicas de aprendizagem desequilibrada e, inversamente, nem todos os métodos usados para lidar com a aprendizagem desequilibrada são apropriados para a aprendizagem sensível ao custo.

A maioria dos classificadores assume que os custos de classificação incorreta (custo falso negativo e falso positivo) são os mesmos. Na maioria dos aplicativos do mundo real, essa suposição não é verdadeira [63].

Tradicionalmente, os algoritmos de aprendizado de máquina são treinados em um conjunto de dados com o objetivo de minimizar erro de classificação. Nesse contexto, ajustar um modelo aos dados resolve um problema de otimização, em que procuramos explicitamente minimizar o erro na classificação das observações. A minimização do erro, custo ou perda é o que se quer quando um algoritmo é submetido ao treinamento nos dados. A aprendizagem sensível ao custo é um tipo de aprendizagem que leva em consideração os custos da classificação incorreta, tendo como objetivo minimizar o custo total.

Segundo Peter [64], há algumas dimensões de custo, de aplicações do mundo real, que podem ser consideradas no processo de desenvolvimento dos modelos. A maior parte da literatura sobre aprendizado de máquina ignora todos os tipos de custo. O autor exemplifica alguns custos que podem ser ponderados na modelagem:

i. Custo de erros de classificação (ou erros de previsão de forma mais geral).

ii. Custo de testes ou avaliação.

iii. Custo de rotulagem.

iv. Custo da intervenção ou alteração do sistema fonte dos dados utilizados no treinamento.

v. Custo de realizações ou resultados indesejados da intervenção.

vi. Custo de computação ou complexidade computacional.

vii. Custo dos casos ou coleta de dados.

viii. Custo da interação humano-computador ou enquadramento do problema e uso de software para ajustar e usar um modelo.

ix. Custo de instabilidade ou variação conhecido como desvio de conceito.

Os métodos de aprendizado de máquina sensíveis ao custo são aqueles que usam explicitamente a matriz de custos. Dado nosso foco na classificação desequilibrada, estamos especialmente interessados nas técnicas sensíveis ao custo que se concentram no uso de custos variáveis de classificação incorreta de alguma forma.

Dentre as abordagens disponíveis, são destacados três grupos mais relevantes para aprendizado desbalanceado [59]:

a) Reamostragem sensível ao custo: A amostragem de dados é uma técnica que pode ser adotada diretamente, com custo reduzido. Em vez de amostrar com foco no equilíbrio da distribuição desbalanceada das classes, o foco está na alteração da composição do conjunto de dados de treinamento para atender às expectativas da matriz de custo. Isso pode envolver a amostragem direta da distribuição de dados ou o uso de um método para ponderar exemplos no conjunto de dados. Tais métodos podem ser referidos como ponderação proporcional ao custo do conjunto de dados de treinamento ou amostragem proporcional ao custo.

b) Algoritmos sensíveis ao custo: Algoritmos de aprendizado de máquina raramente são desenvolvidos especificamente para aprendizado sensível ao custo. Mas, eles podem ser modificados para fazer uso da matriz de custo. Muitas dessas modificações foram implementadas para algoritmos populares, como árvores de decisão e máquinas de vetores de suporte.

c) *Ensembles* sensíveis ao custo: Grupo distinto de métodos, com técnicas projetadas para filtrar ou combinar as previsões dos modelos tradicionais de aprendizado de máquina, a fim de levar em conta os custos de classificação incorreta. Esses métodos são chamados de métodos de "*wrapper*", pois envolvem um classificador de aprendizado de máquina padrão. Eles também são chamados de *meta-learners* ou *ensembles*, pois aprendem como usar ou combinar previsões de outros modelos.

O *meta-learning* sensível ao custo converte algoritmos classificadores insensíveis ao custo, em classificadores sensíveis ao custo, sem modificar estrutura, parâmetros ou hiper parâmetros. Assim, pode ser considerado um componente de *middleware* que pré-

processa os dados de treinamento, ou pós-processa a saída, a partir dos algoritmos de aprendizagem insensíveis ao custo.

2.2.5. Algoritmos utilizados em classificação do risco de crédito

O campo de pesquisa da Mineração de Dados envolve o uso de ferramentas computacionais, como algoritmos de *machine learning* ou aprendizado de máquina, para auxiliar na solução dos mais variados problemas. Segundo Leo, Sharma e Maddulety [65] essa área incorpora aspectos da ciência da computação, engenharia e estatística, destacando-se como um meio comumente aplicado a problemas reais, sobretudo quando se faz necessário interpretar dados. Há que se destacar que o aprendizado de máquina possui no nome a palavra "aprendizado" por conta da capacidade de dotar os algoritmos ou modelos de aprender ou de se adaptar, a partir do treinamento em um conjunto de dados.

No âmbito da Mineração de Dados são aplicadas também técnicas de aprendizado supervisionado e não supervisionado para classificação e regressão, no contexto de serviços financeiros essas ferramentas são as mais comumente utilizadas, junto com sistemas de otimização inteligentes [66]. Quando há problemas em que se tem informações ou dados que vão ser utilizados (treinados) para prever algo observável (*target*), este tipo de abordagem é chamada de supervisionado. Já quando não há uma *target* e o que será realizado está no campo de descrever associações e padrões a partir de um conjunto de dados, tal abordagem é chamada de não supervisionada.

Em trabalho de revisão de literatura Barboza [48] destaca que embora o poder preditivo de modelos baseados em estatísticas multivariadas tenha diminuído, as técnicas mais recentes de inteligência artificial e aprendizado de máquina representam uma nova linha de pesquisa para risco de crédito, especificamente para previsão de falências.

Na literatura vários pesquisadores [65][67][68][69] destacam o uso de técnicas de *machine learning* como alternativas preditivas melhores do que as tradicionalmente utilizadas nos processos para prever o risco em instituições financeiras, como as técnicas estatísticas multivariadas mais conhecidas: Análise Discriminante e Regressão Logística. Dentre os algoritmos de *machine learning* que se destacam na modelagem de *Credit Scoring*, com classes desbalanceadas, estão o *Random Forest* e o *XGBoost* [69].

No contexto de avaliar a capacidade de pagamento ou a dificuldade financeira de entes públicos, como municípios, existem pesquisas que utilizam diferentes abordagens e métodos de análise, como técnicas estatísticas de regressão logística e análise discriminante e, mais recentemente, métodos baseados em abordagens heurísticas, como os métodos multicritério [17][18][20][21].

Os algoritmos estão disponíveis nas mais variadas linguagens, particularmente em *Python* a biblioteca mais utilizada é a *scikit-learn*[5] e a *XGBoost*[6]. A seguir estão detalhadas as técnicas mais comumente utilizadas em processos de avaliação do risco de crédito, conforme comentado anteriormente.

Regressão Logística

A Regressão Logística (RL) é um classificador que possui a capacidade de gerar cálculos probabilísticos, com estimativa de máxima verossimilhança, o que significa que suas probabilidades já estão calibradas, diferentemente de outros classificadores de *machine learning* [59].

Essa técnica estatística é comumente utilizada no desenvolvimento de modelos de avaliação do risco de crédito, porque a variável dependente (*target*) pode ser reduzida a um resultado binário ([0,1]; "bons" ou "ruins"), em que estimativa probabilística de se chegar no resultado (0,1) vem da aplicação do algoritmo em um conjunto de dados (variáveis preditoras), o processo acontece por meio da estimativa de máxima verossimilhança. A RL estima a probabilidade de que um indivíduo com determinado perfil "p_i" seja um cliente "bom", por exemplo, essa probabilidade pode ser denotada por $p(\text{"}bom\text{"}|k)$. O modelo logístico está fundamentado na validade da relação:

$$logit\ (p_i) = ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{2.1}$$

Denotando a função linear por Z, temos:

$$Z = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{2.2}$$

Assim:

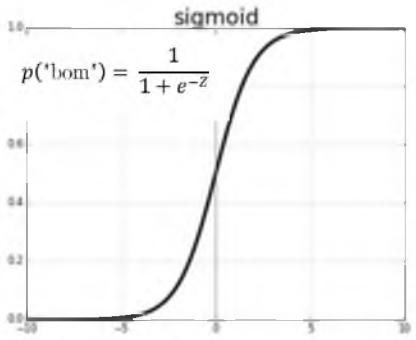
$$Z = ln\left(\frac{p_i}{1-p_i}\right) \tag{2.3}$$

Portanto, sendo i = "bom", temos:

$$p(\text{"bom"}) = \frac{e^Z}{1-e^Z} = \frac{1}{1+e^{-Z}} \tag{2.4}$$

Graficamente a função Logística ou a variação de p("bom") com Z pode ser vista na Figura 2.10 a seguir:

[5] https://scikit-learn.org/stable/index.html
[6] https://xgboost.readthedocs.io/en/latest/



Fonte: Sicsú [70] (adaptado)

Figura 2.10: *Representação da função logística (sgmoid)*

Random Forest

O algoritmo *Random Forest* (RF) foi desenvolvido por Breiman [71] e apresentado em 2001, em termos estruturais a sua construção é bem diferente do que observamos na Regressão Logística. O RF é considerado um método *ensamble* de aprendizado em que são construídos vários "modelos", no caso árvores de decisão, empacotados (*bagging*), a partir e seleção de amostra de *bootstrap* do conjunto de dados de treinamento. No processo do RF, nem todas as variáveis ou observações são utilizadas, sendo escolhido um pequeno conjunto de variáveis para cada amostra de *bootstrap*. Essa operação tem efeito de descorrelacionar as árvores de decisão, assim cada árvore será diferente das demais, tornando-as mais independentes. O processo é repetido n vezes, sem "podar" as árvores, deixando crescer de forma otimizada, por conseguinte melhorando a previsão do conjunto.

O algoritmo pode ser ajustado para buscar o melhor resultado, considerando as características dos dados e o que se quer medir, por meio de parâmetros, como por exemplo: nº de árvores da floresta; critério para medir a qualidade de uma divisão da árvore (ex.: gini, entropia); profundidade máxima da árvore; nº mínimo de amostras; nº mínimo de folhas; quantidade máxima de variáveis ou observações, calculada pela raiz quadrada ou log da quantidade; máximo de nós das folhas; peso para ponderar as classes, entre outros.

Graficamente um modelo com base no RF tem uma seguinte estrutura:

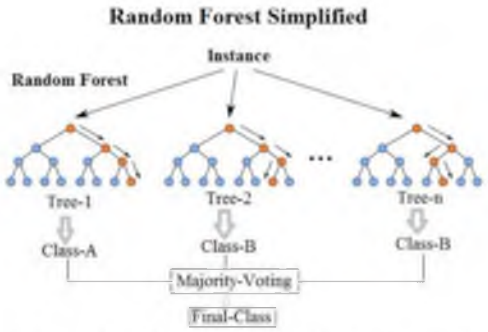


Figura 2.11: *Representação do Random Forest*[7]

No entanto, semelhante à maioria dos classificadores de *machine learning*, RF também pode sofrer com a "maldição" de aprender com um conjunto de dados de treinamento extremamente desequilibrado, conforme observado por Breiman [72] em seu trabalho de 2004. Nesse artigo, são propostas duas maneiras para lidar com o problema de classificação de dados desequilibrados: (i) aprendizado sensível ao custo e (ii) aprendizado baseado em técnica de amostragem.

XGBoost

A palavra *XGBoost* vem de *Extreme Gradient Boosting*, esse algoritmo de *machine learning* foi criado por Chen e Guestri [73], sendo também baseado em árvores de decisão, como o *Random Forest*, e utiliza uma estrutura de gradiente *boosting*[8]. No processo de geração, o algoritmo cria um modelo encadeado a partir de árvores geradas de forma independente (diferentes). São criadas árvores pequenas, com baixo viés e alta variabilidade (variância), combinando dessa forma vários modelos "fracos" que ao final se chega a um modelo otimizado. Nesse processo, os resíduos gerados pelas árvores são utilizados como observações para modelagem de novas árvores e assim sucessivamente, até chegar em um modelo com no mínimo de erro. A Figura 2.12 mostra graficamente um exemplo final da função de um modelo gerado por meio do *XGBoost*.

[7] Fonte: https://community.tibco.com/wiki/random-forest-template-tibco-spotfire

[8] *Boosting* pode ser interpretado como um algoritmo de otimização de uma função de custo adequada, sendo o gradiente *boosting* uma técnica de *boost* em que, para cada iteração, um novo preditor é construído para se ajustar aos pseudo-resíduos do preditor anterior.

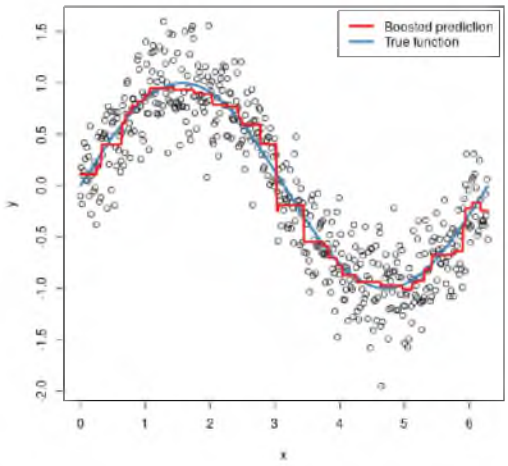


Figura 2.12: *XGBoost – exemplo de uma função preditiva*[9]

O *XGBoost* dispõe de um conjunto amplo de hiperparâmetros, possibilitando controle refinado no processo de treinamento do modelo, por exemplo: tipo de modelo a ser executado em cada interação (ex.: árvore ou linear); nº de árvores; limitação de *boosting*; taxa de aprendizado do modelo; definição de pesos das variáveis; profundidade das árvores; taxa de amostragem; parâmetros de regularização (reduzir complexidade do modelo); peso das classes; definição da função objetivo do modelo, entre outros.

O algoritmo *XGBoost* é eficaz para uma ampla gama de problemas de modelagem preditiva de regressão e classificação. Embora o algoritmo tenha um bom desempenho em geral, mesmo em conjuntos de dados de classificação desequilibrados, ele oferece uma maneira de ajustar o algoritmo de treinamento para prestar mais atenção à classificação incorreta da classe minoritária para conjuntos de dados com uma distribuição de classe distorcida [59].

2.2.6. Métricas de avaliação

As medidas de avaliação da performance dos modelos se apresentam fundamentais na avaliação do desempenho da classificação e na orientação da modelagem do algoritmo classificação, em problemas de classificação relacionados a bases fortemente desbalanceadas [54].

Brownlee [59] destaca que a seleção de um modelo e até mesmo dos métodos de preparação de dados (amostragem) é um problema de pesquisa que é orientado pela métrica de avaliação, em que os experimentos são realizados com algoritmos diferentes e o resultado de cada modelo é medido pelos indicadores de performance.

[9] Fonte: https://sigmoidal.ai/xgboost-aprenda-algoritmo-de-machine-learning-em-python/

Vários pesquisadores de aprendizado de máquina identificaram três famílias de métricas de avaliação dentre as medidas comumente utilizadas em avaliação dos modelos de classificação, desenvolvidos em mineração de dados: (i) métricas de limites, por exemplo, Acurácia e *F-mensure*; (ii) métrica de ranking, por exemplo a Curva ROC, e (iii) métricas de probabilidade, ex.: erro quadrático médio. Considerando o efeito de classes desbalanceadas serão destacadas as medidas listadas a seguir [59].

Medidas de limite

A métrica provavelmente mais utilizada é a acurácia. Para classes desbalanceadas, essa métrica não se apresenta como a mais recomendada pois pode trazer uma falsa informação de que o modelo tem boa performance já que está acertando bem os eventos na classe majoritária, contudo o modelo pode não prever as observações na classe minoritária.

$$Acurácia = \frac{Previsões\ corretas}{Total\ de\ previsões} \tag{2.5}$$

Em complemento à Acurácia, temos:

$$Erro = \frac{Previsões\ incorretas}{Total\ de\ previsões} \tag{2.6}$$

A partir das classificações corretas e incorretas é possível montar uma matriz das medidas calculadas versus o que a base de informações traz. Essa matriz é chamada de Matriz de Confusão, graficamente ela é assim representada:

Quadro 2.2: *Matriz de confusão binária*

		Previsão do modelo	
		Previsão Positiva	**Previsão Negativa**
Observado	**Classe Positiva**	Verdadeiro Positivo (TP)	Falso Negativo (FN) Erro II
	Classe Negativa	Falso Positivo (FP) Erro I	Verdadeiro Negativo (TP)

Fonte: Autoria própria

Outro grupo de métricas que são úteis em situação de classes desbalanceadas, dada a característica de focar em uma classe, são a sensibilidade, a especificidade, a precisão e o recall.

A Sensibilidade é uma taxa para avaliar o quanto o modelo acertou a classe positiva:

$$Sensibilidade = \frac{TP}{TP + FN} \tag{2.7}$$

A Especificidade é a taxa de para avaliar o quanto o modelo acertou a classe negativa:

$$Especificidade = \frac{TN}{FP + TN} \tag{2.8}$$

Situações como a estudada neste trabalho, o desejável é que o modelo consiga classificar corretamente as duas classes, já que no contexto da gestão dos riscos e negócios há perdas potenciais para a instituição nas duas situações de erro tipo I e II.

A Sensibilidade e a Especificidade podem ser combinadas em uma única pontuação que equilibra as duas preocupações, chamada de média G.

$$G - mean = \sqrt{Sensibilidade \times Especificidade} \tag{2.9}$$

A Precisão verifica a frequência de acerto dentre aqueles classificados no evento de interesse:

$$Precisão = \frac{TP}{TP + FP} \tag{2.10}$$

O *Recall* tem a mesma forma de cálculo da Sensibilidade.

$$Recall = \frac{TP}{TP + FN} \tag{2.11}$$

As métricas Precisão e Recall podem ser combinadas em um único indicador que equilibra os dois conceitos, trazendo um valor que indique a qualidade geral do modelo, sendo chamada de *F-score* ou *F-measure*.

$$F - measure = \frac{2 \times Precisão \times Recall}{Precisão \times Recall} \tag{2.12}$$

Medidas de *Ranking*

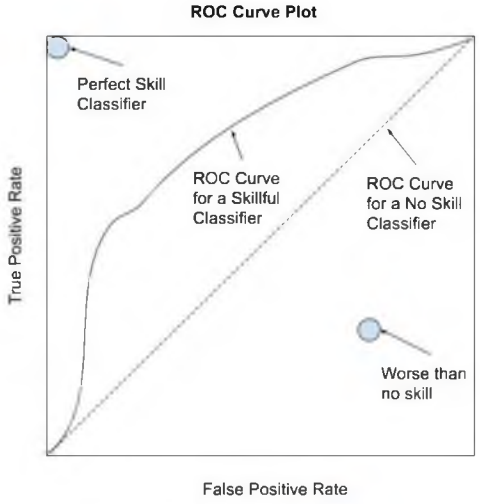
Na literatura [59][74], a métrica de *ranking* mais comumente usada é a Curva ROC ou Análise ROC. ROC é uma sigla que significa *Receiver Operating Characteristic*, trazendo um campo de estudo para analisar classificadores binários com base em sua capacidade de discriminar classes. A Curva ROC é um gráfico de diagnóstico que resume o comportamento de um modelo, calculando a taxa de falso positivo e a taxa de verdadeiro positivo para um conjunto de predições do modelo sob diferentes limites. A verdadeira taxa positiva é o *Recall* ou Sensibilidade.

$$Taxa\ de\ verdadeiros\ positivos\ (TPR) = \frac{TP}{TP + FN} \tag{2.13}$$

A taxa de falsos positivos é calculada da seguinte forma:

$$Taxa\ de\ falsos\ positivos\ (FPR) = \frac{FP}{FP + TN} \tag{2.14}$$

Graficamente a Curva ROC mostra a habilidade de o modelo classificar as observações nas classes, o resultado quanto mais próximo do quadrante esquerdo no alto melhor. A Figura 2.13 traz uma visualmente como interpretar a Curva ROC.



Fonte: Brownlee [59]

Figura 2.13: *Curva ROC – exemplo*

Cabe destacar que a Curva ROC é colocada pelos pesquisadores como eficaz, mas medidas como AUC e Curva ROC podem ser otimistas sob um cenário de desequilíbrio de classe severo, especialmente quando o número de exemplos na classe minoritária é pequeno.

Uma medida popular no contexto de avaliação dos modelos de risco de crédito é a estatística *Kolmogorv-Smirnov* (KS) [70][75]. O KS é um teste não-paramétrico para comparar duas amostras de uma mesma população, para avalia a separação relativa entre "bons" e "ruins", com o objetivo é inferir se são distintas. Quanto maior o valor, maior é a diferença entre os dois grupos, assim é possível verificar se o modelo está cumprindo seu objetivo.

O cálculo é realizado a partir da maior diferença entre as distribuições acumuladas das probabilidades ou escores dos "bons" $F_b(k)$ e "ruins" $F_r(k)$.

$$F_b\ (k) = \frac{números\ de\ bons\ com\ escore\ de\ risco\ \leq k}{número\ de\ bons} \tag{2.15}$$

$$F_r\ (k) = \frac{números\ de\ ruins\ com\ escore\ de\ risco\ \leq k}{número\ de\ ruins} \tag{2.16}$$

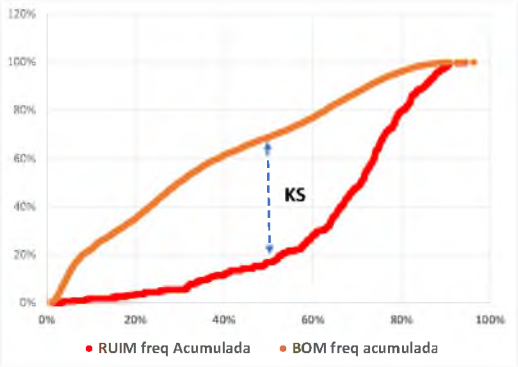
Onde:

k = conjunto de possíveis valores de escore de risco.

Calcula-se os valores das funções (2.15) e (2.16, para k variando dentro da amplitude de valores da escala de risco. Na sequência, é calculado o valor do KS a partir da maior diferença entre as duas funções.

$$KS = máx\,[F_r(k) - F_b(k)] \tag{2.17}$$

Graficamente as distribuições acumuladas de "bons" e "ruins" e o respectivo KS é representado como na Figura 2.14.



Fonte: Autoria própria

Figura 2.14: *Curva KS – exemplo*

3. Método de Pesquisa

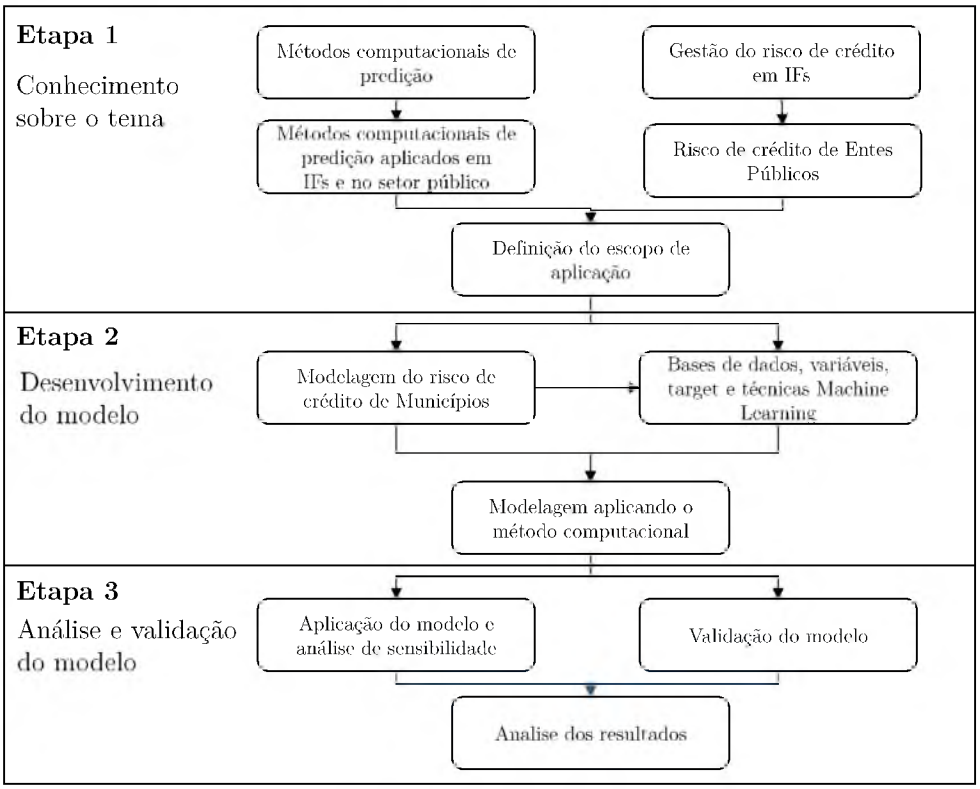
O método de pesquisa consiste na sequência de passos necessários para demonstrar que o objetivo proposto na pesquisa foi atingido, segundo Wazlawick [76]. A seguir, são listados os passos do desenvolvimento na pesquisa, os meios adotados e a os resultados observados.

3.1. Tipo de pesquisa

Conforme Wazlawick [77], o tipo de pesquisa em que se enquadra este trabalho é o tipo "Exploratória", por ser de natureza aplicada, com abordagem quantitativa e empregar técnicas específicas para tratar problemas observados no mundo real.

3.2. Metodologia

O presente estudo propõe-se a apresentar uma forma de avaliação e classificação de ente públicos (municípios), no contexto de gestão de portfólios em Instituições Financeiras. A pesquisa está estruturada em três etapas de desenvolvimento, tendo como referência o processo de tomada de decisão apresentado por Belton & Stewart [78]. As etapas do método de pesquisa estão apresentadas na Figura 3.1.



Fonte: Autoria própria

Figura 3.1: *Etapas do Método de Pesquisa*

Na primeira etapa é contextualizado o ambiente interno e externo em que as Instituições Financeiras estão inseridas na gestão de portfólios de entes públicos, trazendo a identificação do problema de gerenciamento de portfólio de crédito em IFs. A pesquisa bibliográfica destacou as principais referências, relacionadas à gestão de riscos de créditos em IFs e à mensuração do risco de crédito utilizando métodos computacionais e os tradicionalmente adotados nos processos de modelagem. Essas etapas precedem a definição do escopo de aplicação dos referenciais teóricos, no contexto do problema identificado vis-à-vis o objetivo da pesquisa.

Na segunda etapa, são estruturadas as bases de dados que serão utilizadas para construção da base de modelagem, com a definição da variável *target* (risco de crédito) e as variáveis explicativas, a partir dos dados coletados em fontes abertas na internet, em sites oficiais, complementadas por dados disponibilizados pela IF de grande porte. Essas variáveis passam por um processo de avalição qualitativa e quantitativa, em que foram observadas as características dos municípios. Tais variáveis trazem dados particulares e diversificados sobre cada ente público e possuem as mais variadas informações, como por exemplo: nível de gestão pública; desenvolvimento socioeconômico; características regionais; aspectos demográficos; regimes previdenciários, entre outros. Outro ponto observado na construção da base de modelagem, foi o nível de atualização dos valores das variáveis no tempo, para que seja possível capturar o dinamismo do desenvolvimento dos municípios e a relação com a variável *target*. Nessa etapa também são definidos os métodos computacionais para aplicação no problema de predição do risco de crédito, partindo para o processo de modelagem propriamente dito, em que são utilizadas técnicas para tratamento do desbalanceamento das classes (*low default*) e uso de indicadores de performance para verificar o nível de acurácia dos modelos desenvolvidos.

Por fim, na terceira etapa é realizada a validação dos modelos desenvolvidos na etapa anterior, aplicando-os em uma base de dados (*out-of-time*) que não foi utilizada na fase de modelagem como treinamento. Nesse momento, observa-se o comportamento dos modelos numa base "desconhecida" por eles, sendo avaliada a performance, procurando verificar se os modelos possuem comportamento de *overfitting* ou *underfitting*. Os modelos são analisados em termos de indicadores, matriz de confusão, curva ROC entre outras medidas, sendo comparados de modo a verificar quais entre eles apresentam os melhores resultados e se há diferença significativa, a partir da medida de acurácia. Além da análise dos modelos "por dentro", foi realizada uma comparação entre os resultados dos modelos com a avaliação de risco gerada pelo Ministério da Economia, por meio da CAPAG. Por fim, é realizada a conclusão do estudo com destaque para os resultados observados e sugestão de futuros estudos.

3.3. Objeto de estudo

O objeto de estudo são os 5.570[10] municípios brasileiros e as suas características disponíveis em bases de dados abertas, dados econômico-financeiros, sociais, características regionais, descritas com mais detalhe no Apêndice.

3.4. Fonte dos dados e horizonte disponível

O processo utilizado por pesquisadores para levantar variáveis explicativas, relacionadas a eventos de dificuldade financeira ou de capacidade de pagamento (*default*), direcionam para o uso de informações relacionadas à população (tamanho populacional), às características socioeconômicas, à saúde financeira, fiscal e orçamentária. Observa-se o uso dessas informações em estudos ao redor do mundo, particularmente na Europa (Grécia, Espanha e França) [17][18][19][21] e nos EUA [20]. Embora alguns desses trabalhos não tenham examinado municípios de pequeno e médio porte, seus achados para os entes públicos maiores justificam a seleção de variáveis explicativas semelhantes, que serão utilizadas no presente estudo.

Quanto ao uso de indicadores sintéticos socioeconômicos, relacionados a entes públicos, Guimarães e Jannuzzi [79], em estudo sobre políticas públicas, reconhecem a importância desses indicadores nas esferas técnicas e políticas no país como referência para processos relacionados às decisões sobre políticas públicas nos últimos anos. Todavia, são destacadas inadequações de uso desses indicadores, em processos de tomada de decisão com base em um único referencial, por não levar em conta as diversas limitações metodológicas e conceituais desses índices.

A partir das referências utilizadas em trabalhos semelhantes[3][17][20], foi realizada pesquisa na internet em fontes de dados abertos, de origem governamental ou de órgãos de pesquisa e estudos regionais, contendo informações dos municípios brasileiros, nas dimensões as mais variadas possíveis, de modo a capturar as nuances socioeconômicas, demográficas, fiscais e regionais.

As fontes de dados encontradas foram as seguintes:

a) **Dados abertos no Atlas da Vulnerabilidade Social**

[10] https://cidades.ibge.gov.br/brasil/panorama

Gerado pelo Instituto de Pesquisa Econômica Aplicada (IPEA) [80]. Um compilado de vários atributos e índices sociais em uma base de dados com 101 campos (Apêndice – Quadro 8.1). Entre as informações disponíveis, destacam-se:

i. O Índice de Vulnerabilidade Social (IVS), um índice sintético que agrega um conjunto de variáveis quantitativas, cuja fonte é o censo do IBGE. Este índice é composto por 16 indicadores relacionados a situações de vulnerabilidade social, com detalhamento em várias dimensões chegando ao nível dos municípios brasileiros;

ii. Índice de Desenvolvimento Humano Municipal (IDHM), também um índice sintético desenvolvido a partir dos mesmos conceitos e dimensões do Índice de Desenvolvimento Humano Global (IDH Global), composto pelas visões de longevidade, educação e renda;

iii. A combinação desses dois indicadores, o indicador de "Prosperidade Social", que apresenta uma visão sobre como cada ente público está posicionado simultaneamente, quanto aos níveis de desenvolvimento humano e de vulnerabilidade social, por exemplo quando um município possui um alto nível de desenvolvimento humano e uma baixa vulnerabilidade social, observa-se uma trajetória de desenvolvimento humano menos vulnerável e socialmente mais próspera. Essa base de dados possui pouca dinâmica, as informações mais atualizadas são referenciadas no último censo demográfico brasileiro de 2010.

b) **Dados abertos sobre a atividade econômica dos municípios:**

Informações disponibilizadas pelo Instituto Brasileiro de Geografia e Estatística (IBGE) contém um conjunto de 43 campos estruturados, construídos na periodicidade anual e histórico de 2010 a 2017 (Apêndice – Quadro 8.2), com destaque para: Produto Interno Bruto (PIB); Valor Adicionado Bruto a partir da atividade econômica (4 níveis); característica regional (semiárido ou não, rural ou não), entre outras informações.

c) Dados abertos disponibilizados pela Secretaria do Tesouro, vinculada ao Ministério da Economia

Destaque para a métrica de análise da capacidade de pagamento (CAPAG) para apurar a situação fiscal dos entes públicos (estados e municípios). O intuito da CAPAG é apresentar de forma simples e transparente se um novo endividamento representa risco de crédito para o Tesouro Nacional.

A metodologia do cálculo, dada pela Portaria MF nº 501/2017 [81], é composta por três indicadores: endividamento, poupança corrente e índice de liquidez. Na

metodologia, o critério mais relevante é o de Liquidez, seguido do Indicador de Poupança Corrente e do Endividamento. Avaliando o grau de solvência, a relação entre receitas e despesa correntes e a situação de caixa é realizado um diagnóstico da saúde fiscal do estado ou município. Os conceitos e variáveis utilizadas e os procedimentos a serem adotados na análise da CAPAG foram definidos na Portaria STN nº 882/2018 [82].

Consultando o site do Ministério da Economia, Secretaria do Tesouro Nacional[11], observa-se que as bases de dados disponíveis, em formato ".csv", trazem informação da classificação da CAPAG dos municípios, apurados quadrimestralmente, mas não há disponibilidade de dados históricos. Na data-base de abril de 2021, por exemplo, verificou-se que nem todos os municípios possuem "*rating*" na visão da CAPAG. Entre os 5.570 municípios, 20,7% não possuíam as classificações possíveis nos níveis "A", "B", "C" ou "D" e receberam a identificação "n.d.", significando que o município não possui uma classificação CAPAG.

Esse percentual de não classificados é elevado, correspondendo a 1.151 municípios, portanto nesse contexto não seria possível utilizar a métrica e nem os três indicadores que a compõem. Outro ponto importante, conforme as informações da metodologia da CAPAG [81], a informação dos níveis de classificação não fazem menção às estimativas de probabilidade de pagamento ou de *default*, não sendo portanto uma metodologia a ser adotada para estimar perdas esperadas de crédito.

d) Dados abertos disponibilizados pela Federação das Indústrias do Estado do Rio de Janeiro (FIRJAN):

São informações anuais sobre os municípios brasileiros, com histórico de 2013 a 2018 (Apêndice – Quadro 8.3). Destaque para o Índice Firjan Gestão Fiscal (IFGF) [83], o qual é inteiramente construído com base em resultados fiscais oficiais, declarados pelas próprias prefeituras. Conforme estabelecido pelo Artigo 51, da Lei de Responsabilidade Fiscal [9], os municípios devem encaminhar suas contas para a Secretaria do Tesouro Nacional (STN), até o dia 30 de abril do ano seguinte ao exercício de referência, a partir de quando o órgão dispõe de 60 dias para disponibilizá-las ao público, por meio do Sistema de Informações Contábeis e Fiscais do Setor Público Brasileiro (Siconfi). Segundo a documentação da Metodologia [55], esse instrumento consolida informações contábeis, financeiras e estatísticas fiscais oriundas de um universo que compreende 5.568 municípios[12], 26 Estados, o Distrito Federal e a União. O Siconfi é a principal fonte de dados sobre as administrações públicas municipais e estaduais. Por isso, foi utilizado

[11] http://www.tesourotransparente.gov.br/ckan/dataset/capag-municipios

[12] Brasília e Fernando de Noronha não entraram na base de dados por não possuírem prefeitura

como referência para o cálculo do IFGF. A despeito da determinação legal, na divulgação dos dados referentes ao exercício fiscal de 2018 não constavam na base de dados do Siconfi as informações necessárias para o cálculo do índice de 100 municípios[13]. Conforme informado no relatório, no processo de tratamento dos dados foram descartados os dados de 131 municípios por apresentarem inconsistências que impediram a análise. Logo, o cálculo do IFGF ano-base 2018 foi possível para 5.337 municípios, onde vivem 200,9 milhões de pessoas – 97,8% da população brasileira.

O IFGF tem como objetivo contribuir com o debate sobre a eficiência da gestão fiscal, com foco na administração dos recursos públicos efetuado pelas prefeituras brasileiras. Em 2019, o IFGF passou por reformulação e o novo índice passou a ser composto por quatro indicadores, que assumem o peso de 25% na composição do índice geral, são eles:

i. IFGF Autonomia – evidencia um dos pontos mais críticos para a gestão fiscal eficiente das prefeituras: a baixa capacidade de se sustentarem;

ii. IFGF Gastos com Pessoal – a despesa com pessoal é o principal item da despesa do setor público, no caso dos municípios representam metade da Receita Corrente Líquida (RCL), em média;

iii. IFGF Liquidez – verifica se os recursos financeiros são suficientes para fazer frente às despesas que foram postergadas para o ano seguinte.

iv. IFGF Investimentos – mede a parcela dos investimentos nos orçamentos municipais.

As informações o IFGF estão em formado de variável contínua e discretizada, numa visão de pior para melhor. O Quadro 3.1 resume a categorização dos indicadores.

Quadro 3.1: *IFGF – Conceitos e Valores*

#	Descrição	Mínimo	Máximo
5	Gestão de Excelência	0,8	1,0
4	Boa Gestão	0,6	0,8
3	Gestão em Dificuldade	0,4	0,6
2	Gestão Crítica	0	0,4
1	Não possui IFGF	-	-

Fonte: FIRJAN [83]

[13] A data final de consolidação do banco de dados do IFGF foi o dia 14 de julho de 2019.

O nível de cobertura do índice da FIRJAN, conforme informação da Federação em seu site[14], o IFGF alcança 5.337 municípios analisados, correspondendo a 95,82% dos municípios brasileiros, superando as informações da CAPAG disponibilizadas pela Secretaria do Tesouro.

e) Dados abertos disponibilizados pela Secretaria de Previdência Social

Destaque para o Indicador de Situação Previdenciária (ISP-RPPS), instituído pela Portaria MF nº 01, de 03 de janeiro de 2017 [84], que acrescentou o inciso V e o parágrafo único ao art. 30 da Portaria MPS nº 402, de 10 de dezembro de 2008 [85]. A composição, metodologia de aferição e periodicidade do ISP foram aprovadas pela Secretaria de Previdência do Ministério da Fazenda, por meio da Portaria SPREV/MF n° 10, de 08 de setembro 2017 [86].

O Indicador de Situação Previdenciária é calculado com base em três grupos de informações, organizados nos seguintes temas centrais (dimensões): Conformidade, Equilíbrio e Transparência. Cada grupo corresponde a um conjunto de verificações e índices, apurado de acordo com sua respectiva metodologia e fontes de informação. A pontuação do Indicador de Situação Previdenciária varia entre 0 (mínimo) e 1 (máximo) [87]. Os dados sobre a previdência social no formato do Indicador de Situação Previdenciária (ISP) são os mais recentes, a Secretaria de Previdência criou esse indicador em 2018, e só há três anos de informação disponível 2018 a 2020.

Considerando essa limitação, optou-se por utilizar a informação do regime previdenciário adotado pelo município, por ser possível estruturar uma série histórica maior (Apêndice –Quadro 8.4), alinhando com os demais dados dos municípios. Vale ressaltar que no Brasil, o sistema previdenciário público está organizado em dois regimes de previdência, cujos trabalhadores do mercado de trabalho formal são vinculados de forma compulsória:

i. Regime Próprio da Previdência Social (RPPS), estipulado no artigo 40 da Constituição Federal [8], na Lei Federal nº 9.717/1998[88] e nas leis de cada ente que adotou esse regime, cujos benefícios são administrados pela unidade gestora previdenciária de cada um dos entes federativos que o instituíram;

[14] https://www.firjan.com.br/ifgf/

ii. Regime Geral da Previdência Social (RGPS), estabelecido no artigo 201 da Constituição Federal [8] e nas leis federais nº 8.212/1991[89] e 8.213/1991[90], cujos benefícios são geridos pelo INSS.

Destaca-se que os RPPS não podem conceder benefícios distintos daqueles previstos no RGPS. Segundo o Ministério do Trabalho e Previdência, Secretaria de Previdência[15], atualmente, há 2.155 RPPS em operação no país: o da União, que abriga todos os servidores públicos efetivos e os militares federais; os 27 estaduais, que atendem todos os servidores públicos efetivos e os militares de cada um dos estados e do Distrito Federal; e os 2.127 municipais, dentre os quais estão todas as capitais, que reúnem os servidores públicos efetivos dos seus respectivos municípios.

Observa-se que nem todas as cidades brasileiras utilizam o RPPS, tendo em vista não obrigatoriedade, restando RGPS como regime adotado para os seus servidores, que nessa situação tanto o ente público como os servidores contribuem diretamente para o Instituto Nacional do Seguro Social (INSS). A limitação do financiamento dos RPPS apenas às contribuições dos servidores efetivos e do ente empregador os expõe a uma fragilidade muito acentuada quanto à reposição de sua base de contribuição. Por estar assentada em um arranjo de repartição, a redução da base de contribuição (servidores efetivos ativos) acarreta desequilíbrios atuariais, gerando a necessidade de injeções financeiras para garantir o pagamento das despesas com os benefícios previdenciários do sistema (servidores inativos).

f) Dados de uma Instituição Financeira (IF), classificada no Segmento S1[16]

São dados coletados por meio de sites com informações abertas, detalhados no Apêndice (Quadro 8.5). Com relação aos dados da IF, as informações disponibilizadas foram extraídas de fontes públicas, no período de 2010 a 2019, e receberam tratamento interno pela IF, sendo que algumas delas foram transformadas em base "100", numa perspectiva de indicador temporal. As variáveis foram renomeadas para uso no estudo, mas sem guardar relação direta com as segmentações de negócio da IF. Quanto à variável

---

[15] https://www.gov.br/previdencia/pt-br/dados-e-estatisticas/painel-estatistico-da-previdencia/regime-de-previdencia-complementar

[16] As informações desses entes foram coletadas por meio de sites com informações abertas, detalhados a seguir, e dados de uma Instituição Financeira classificada no Segmento S1. A Resolução CMN nº 4.553, de 30 de janeiro de 2017 [100], traz o conceito de IF que se enquadram no Segmento S1: bancos que apresentem porte igual ou superior a 10% do Produto Interno Bruto (PIB) brasileiro ou exerçam atividade internacional relevante, contendo dados históricos em vários níveis.

explicada (*target*) o período de observação é menor, compreendendo safras anuais de 2016 a 2019.

Outra fonte pesquisada, na base do IBGE, foi a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua), que tem por objetivo gerar indicadores para acompanhar as flutuações trimestrais e a evolução, a médio e longo prazos, da força de trabalho e outras informações para subsidiar estudos sobre o desenvolvimento socioeconômico do Brasil.

Conforme a Nota Metodológica da PNAD Contínua [91], a pesquisa é realizada por meio de uma amostra probabilística de domicílios, extraída de uma amostra mestra de setores censitários, de forma a garantir a representatividade dos resultados para os diversos níveis geográficos definidos: Brasil, Grandes Regiões, Unidades da Federação e Regiões Metropolitanas que incluem os municípios das capitais. A cada trimestre, a PNAD Contínua investiga em torno de 211.000 domicílios em aproximadamente 16.000 setores censitários.

Embora seja uma base de dados com informações mais atualizadas, a PNAD Contínua não possui informações sobre dados sociodemográficos de todos os municípios brasileiros identificados, não se enquadrando no escopo do objeto de estudo, que é a totalidade dos municípios.

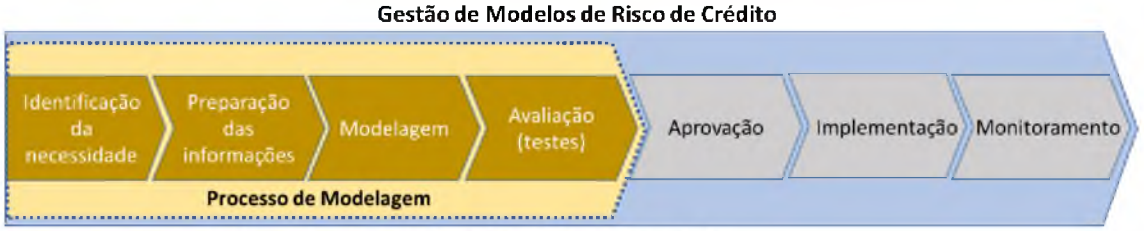
3.5. Processo de modelagem

Segundo Hand e Henley [92], a utilização das técnicas estatísticas para modelagem depende, geralmente, da caracterização detalhada do problema a ser resolvido, da estrutura de dados e dos recursos disponíveis. Nesse contexto, não há que se falar em uma melhor técnica estatística ou matemática a ser aplicada, a melhor decisão quanto à técnica dependerá do propósito da classificação, da disponibilidade das informações e de como serão segregadas (*clusters*) para se chegar ao resultado.

Conforme Sterman [93], a tarefa do modelador é coletar as visões mal definidas e implícitas e montá-las de alguma forma suficientemente bem definida para ser pelo menos entendida e argumentada por outras pessoas, e isso só pode ocorrer se o modelo for totalmente especificado.

O objetivo deste trabalho é explicar a teoria e para isso serão feitas apenas algumas considerações importantes para compreensão geral do processo de modelagem. Isso se deu pois os conhecimentos necessários para a elaboração de modelos, seleção e descrição de variáveis e seus relacionamentos, para a atribuição de valores, definição de relações matemáticas e utilização de softwares e ferramentas de modelagem, exigem prática e estudo.

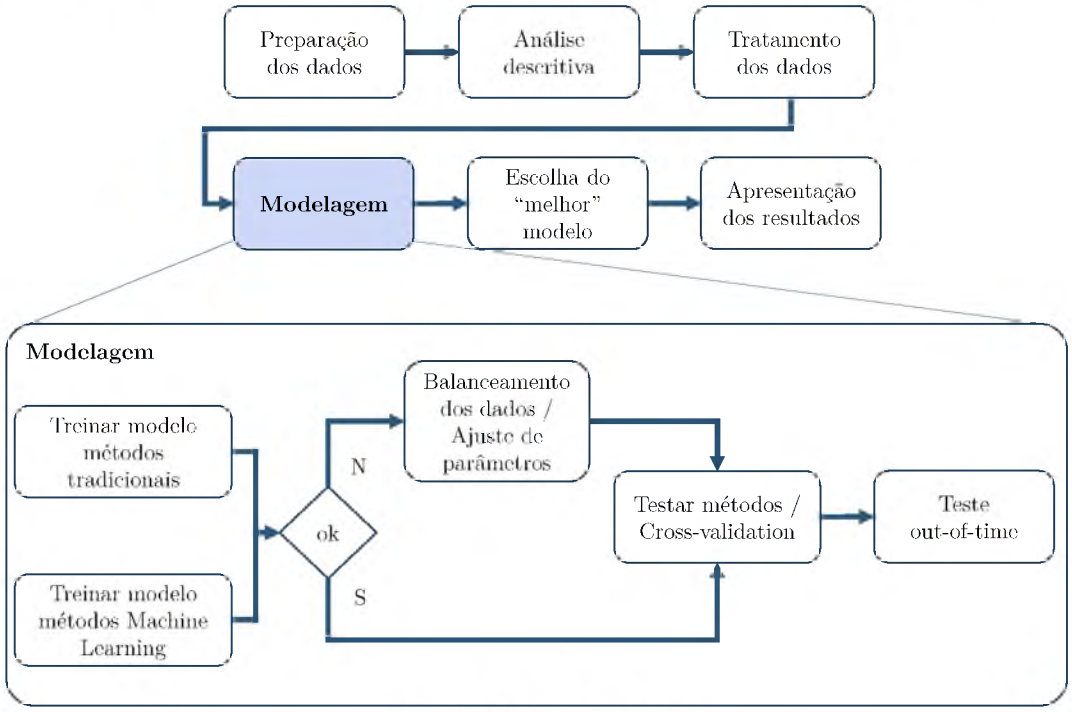
O desenvolvimento dos modelos internos de risco de crédito, normalmente nas IF, obedece a um processo estruturado em várias fases. A Figura 3.2 resume a gestão de modelos de risco de crédito, nela está o processo de modelagem objeto de avaliação dos riscos aplicando a norma ISO 31000:2009 e o modelo CRISP-DM.



(Fonte: Autoria própria)

Figura 3.2: *Processo Gestão de Modelos de Risco de Crédito*

Como foco do estudo, estão destacadas na Figura 3.3 as atividades do processo de modelagem aplicado no desenvolvimento dos modelos, detalhado no Capítulo 4.



Fonte: Autoria própria

Figura 3.3: *Processo de modelagem*

As bases para modelagem de risco de crédito apresentam, por natureza, característica de desbalanceamento das classes. Essa situação de enfatizar demais a classe majoritária e ao mesmo tempo dar atenção insuficiente ao grupo minoritário, segundo S. Lessmann et al. [75] impede a classificação. Tal característica, torna-se um complicador no processo de modelagem, haja visto que o modelo treinado ao ser aplicado nesse contexto de base desproporcional, tende a prever que todo município é "bom", gerando

um alto nível de assertividade das previsões, observado nos indicadores de performance mais adiante.

Para avaliar os efeitos do desbalanceamento, foram construídas também bases de dados com classes balanceadas por meio da amostragem, utilizando técnicas de "*Undersample*" e "*Oversample*", para a estimação dos modelos. Assim, foram avaliados três conjuntos de modelos para a verificar comparativamente as performances: o primeiro utilizando a base de dados original (desbalanceada), o segundo a partir de ajustes nos algoritmos, com foco na sensibilidade ao custo, e o terceiro, utilizando bases de dados balanceadas.

Para avaliação de desempenho foram observadas as seguintes medidas:

a) AUC (Área sob a curva ROC) – Avalia a habilidade do modelo em discriminar os clientes que estão sendo classificados corretamente daqueles que não estão. Uma curva ROC (ou curva de característica de operação do receptor) é um gráfico que resume o desempenho de um modelo de classificação binária na classe positiva. O eixo "x" indica a taxa de falso positivo e o eixo "y" indica a taxa de verdadeiro positivo;

b) *Kolmogorov-Smirnov* (KS) – O KS é um teste não-paramétrico para comparar duas amostras de uma mesma população, para avalia a separação relativa entre "bons" e "ruins", com o objetivo de inferir se são distintas. Quanto maior o valor, maior é a diferença entre os dois grupos, assim é possível verificar que o modelo está cumprindo seu objetivo.

c) Acurácia – Verifica a frequência de acertos do modelo;

d) Precisão – Verifica a frequência de acerto dentre aqueles classificados no evento de interesse;

e) *Recall* – Verifica a capacidade de o modelo identificar corretamente os possuidores do evento de interesse.

f) *F1-score* – Combina as duas métricas anteriores, precisão e *recall*, trazendo um valor que indique a qualidade geral do seu modelo.

3.5.1. Ferramentas de análise

A análise e tratamento dos dados foi realizada por meio dos softwares Microsoft Access e Excel[17] e a Ferramenta Orange [94], Versão 3.27.1. Após tratamento dos dados e construção da base de modelagem, o processo de desenvolvimento dos modelos foi realizado por meio da ferramenta Google Colab[18].

[17] Pacote Microsoft Office 360.

[18] https://colab.research.google.com/

4. Desenvolvimento dos Modelos

4.1. Preparação das informações

4.1.1. Análise descritiva das variáveis e ajustes

Nesta etapa de preparação dos dados inicial, foram utilizadas as ferramentas Microsoft Access, Microsoft Excel e Orange para importar os dados e realizar o pré-processamento, que consiste na análise de relatividade das variáveis, discretização das variáveis de valores contínuas, generalização de dados e eliminação de valores *outliers*, para construção de base de dados que foi utilizada no desenvolvimento dos modelos.

Considerando a grande quantidade de variáveis disponíveis (215, conforme quadros do Apêndice), foi necessário reduzir a lista de dados para o estudo. Nesse sentido, foi realizada uma análise qualitativa dos atributos, em que foi observada a dinâmica das informações, se elas variam ao longo do tempo, informações demográficas (localização, população, atividade econômica principal etc.), e uma análise descritiva para identificar valores faltantes (*missing*) e *outliers*.

O conjunto de dados coletados apresenta detalhes financeiros, econômicos e sociais dos municípios brasileiros. Como o objetivo é determinar se o município é "bom" ou "ruim" sob a ótica de risco de crédito, nesse contexto o processo de modelagem envolve prever se o ente público possui características que permitam avaliar se será ou não um cliente com dificuldade de honrar seus compromissos, por exemplo empréstimos e financiamentos [75].

Após a análise qualitativa dos dados, em que foram excluídos quase todos os valores estáticos gerados com base no Censo demográfico (2010), os valores de variáveis contínuas que possuíam categorização correspondente, os dados correlatos de mesma origem, chegamos no conjunto de dados descritos na Tabela 4.1 a seguir. Essa base contém 21.380 registros, referentes a 4 anos (2015, 2016, 2017 e 2018), com 25 variáveis de entrada, sendo 13 numéricas (inteiras) e 12 categóricas.

Tabela 4.1: *Análise descritiva das variáveis*

#	Variáveis	Qtd.	Média	Desvio Padrão	Mínimo	25%	50%	75%	Máximo
1	rank_IFGF_Autonomia	16351	2.557	1.581	1.000	1.000	2.000	4.000	5.000
2	rank_IFGF_Gastos_com_Pessoal	16351	2.661	1.327	1.000	2.000	2.000	4.000	5.000
3	rank_IFGF_Liquidez	16351	3.018	1.383	1.000	1.000	3.000	4.000	5.000
4	rank_IFGF_Investimentos	16351	2.813	1.125	1.000	2.000	2.000	4.000	5.000
5	regiao	16351	2.914	1.085	1.000	2.000	3.000	4.000	5.000
6	cod_prosp_soc	16351	3.172	1.555	0.000	2.000	3.000	5.000	5.000
7	cod_predomina_pop	16351	1.710	0.454	1.000	1.000	2.000	2.000	2.000
8	cod_semiarido	16351	1.800	0.400	1.000	2.000	2.000	2.000	2.000
9	ativ_econ_maior_vlr_adicionado	16351	2.949	2.387	1.000	1.000	2.000	5.000	10.000
10	pop_faixa	16351	2.417	1.878	1.000	1.000	2.000	3.000	8.000
11	cod_Regime_previdenciario_ajustado_ano	16351	1.390	0.495	1.000	1.000	1.000	2.000	3.000
12	beneficiarios_plan_saude_priv	16351	135.938	57.660	30.000	100.973	122.388	155.672	300.000
13	capitacao_dep_av_ap_SFN	16351	144.999	41.113	30.000	119.065	138.691	161.547	300.000
14	credito_livre_SFN	16351	128.817	48.443	30.000	97.650	121.107	149.268	300.000
15	massa_salarial	16351	150.483	45.533	30.059	122.758	141.039	165.307	300.000
16	estoque_empregos	16351	124.322	31.787	30.035	106.295	118.827	134.969	300.000
17	conexao_banda_larga	16351	5.931	6.681	0.000	1.429	3.699	8.197	100.000
18	diversidade_economica	16351	32.840	24.094	0.000	11.597	29.084	50.952	100.000
19	qtd_empresas	16351	13.307	10.522	0.000	4.300	11.211	20.000	78.581
20	qtd_leitos_hospitalares	16351	5.474	6.746	0.000	0.000	4.154	8.190	100.000
21	qtd_veiculos_leves	16351	31.273	16.057	0.000	17.033	31.280	43.943	100.000
22	qtd_veiculos_pesados	16351	5.521	4.858	0.000	1.815	4.321	8.016	100.000
23	crecimento_populacional	16351	48.598	7.914	0.000	43.712	47.911	52.450	100.000
24	quali_prof_ensino_medio	16351	18.625	10.033	0.000	12.379	16.396	22.330	100.000
25	desc_consolidado	16351	0.038	0.191	0.000	0.000	0.000	0.000	1.000

Fonte: Autoria própria

Para tratamento de "*missing*", decidiu-se por não excluir os registros, dada a escassez de eventos em *default*, sendo criada uma categoria específica para identificar os valores faltantes. Dessa forma, espera-se que os modelos possam classificar os municípios conforme as características originalmente fornecidas, sem necessidade de ajustes.

No conjunto das variáveis escolhidas, há variáveis do tipo contínua, em que para esses atributos se faz necessário realizar a discretização passo importante na modelagem [95], dado que a maioria dos algoritmos de *machine learning* podem ser influenciados, gerando viés no comportamento dos modelos.

No processo de discretização foi adotado método Entropia, segundo H.Liu et al. [95] é uns dos mais utilizados na literatura. A variação do método adotado neste estudo foi disponível na ferramenta Orange [94]: Entropia-MDL. Conforme informação técnica da Orange, esse método utiliza o processo de discretização *top-down*, realizando divisão recursiva do atributo em um corte, em que o ganho de informação seja maximizando até que o ganho seja inferior ao comprimento mínimo de descrição do corte. Esse processo pode resultar em um número arbitrário de intervalos, incluindo intervalo com um valor, caso em que o atributo é descartado (removido).

Os valores dos intervalos discretizados em formato texto foram convertidos em valores numéricos para que fosse possível aplicar o pacote da *scikit-learn* no processo de transformação das variáveis em "*dummy*", citado mais adiante.

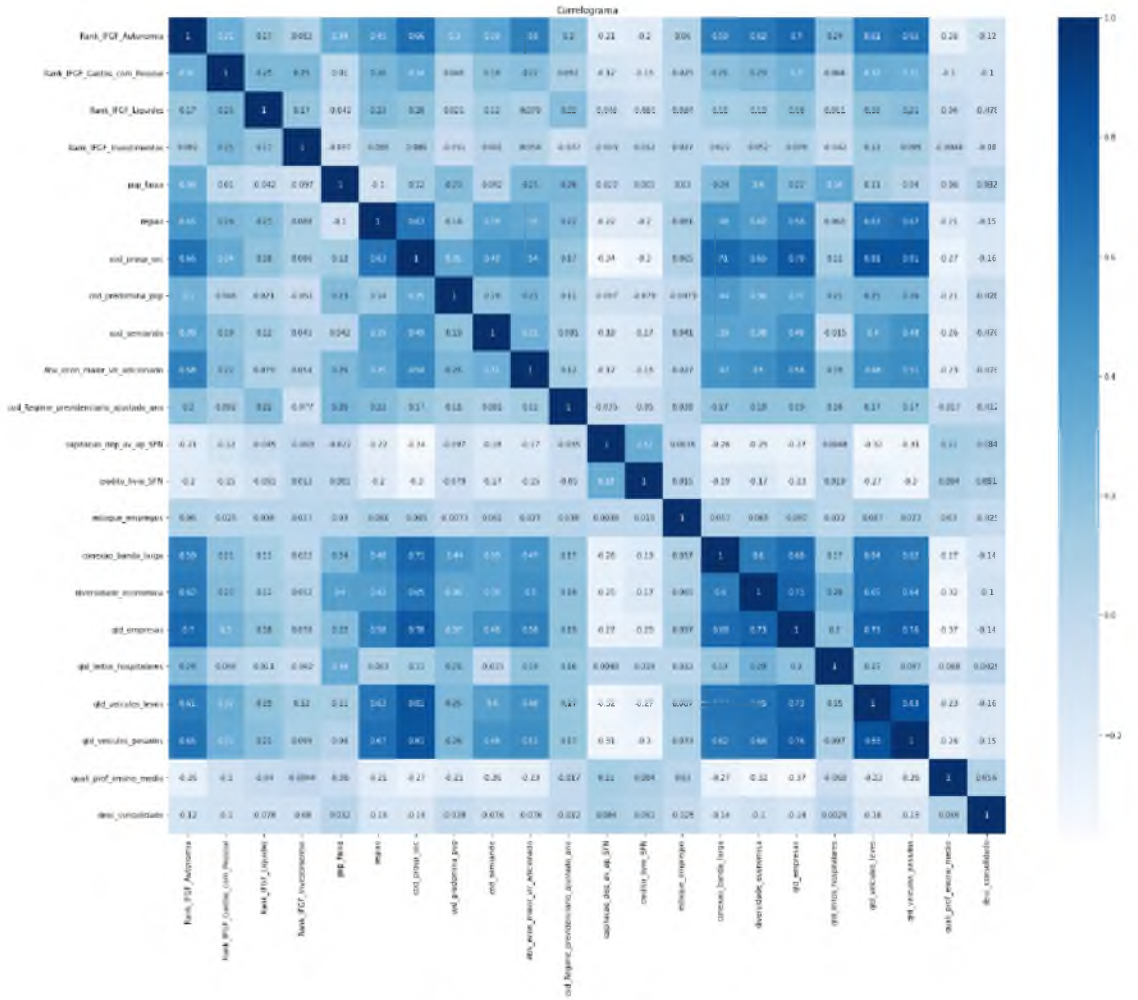
Após o processo de discretização, as variáveis contínuas que permaneceram na base de modelagem, com os respectivos intervalos de classe estão dispostas na Tabela 4.2.

Tabela 4.2: *Conjunto das variáveis discretizadas/categorizadas*

#	Variáveis	Categorias
1	capitacao_dep_av_ap_SFN	< 146.151 146.151 - 180.743 ≥ 180.743
2	credito livre SFN	< 121.779 ≥ 121.779
3	estoque empregos	< 93.4797 93.4797 - 260.741 ≥ 260.741
4	conexao_banda_larga	< 0.879179 0.879179 - 2.77494 2.77494 - 6.73019 ≥ 6.73019
5	diversidade economica	< 11.4311 11.4311 - 24.6384 24.6384 - 43.2446 ≥ 43.2446
6	qtd_empresas	< 6.31734 6.31734 - 10.3283 10.3283 - 27.4584 ≥ 27.4584
7	qtd leitos hospitalares	< 0.106725 0.106725 - 3.55582 3.55582 - 8.35324 ≥ 8.35324
8	qtd_veiculos_leves	< 8.11519 8.11519 - 25.2587 25.2587 - 39.8311 ≥ 39.8311
9	qtd veiculos pesados	< 2.91241 2.91241 - 3.98568 3.98568 - 6.68388 ≥ 6.68388
10	quali_prof_ensino_medio	< 6.23392 6.23392 - 23.7808 ≥ 23.7808

Fonte: Autoria própria

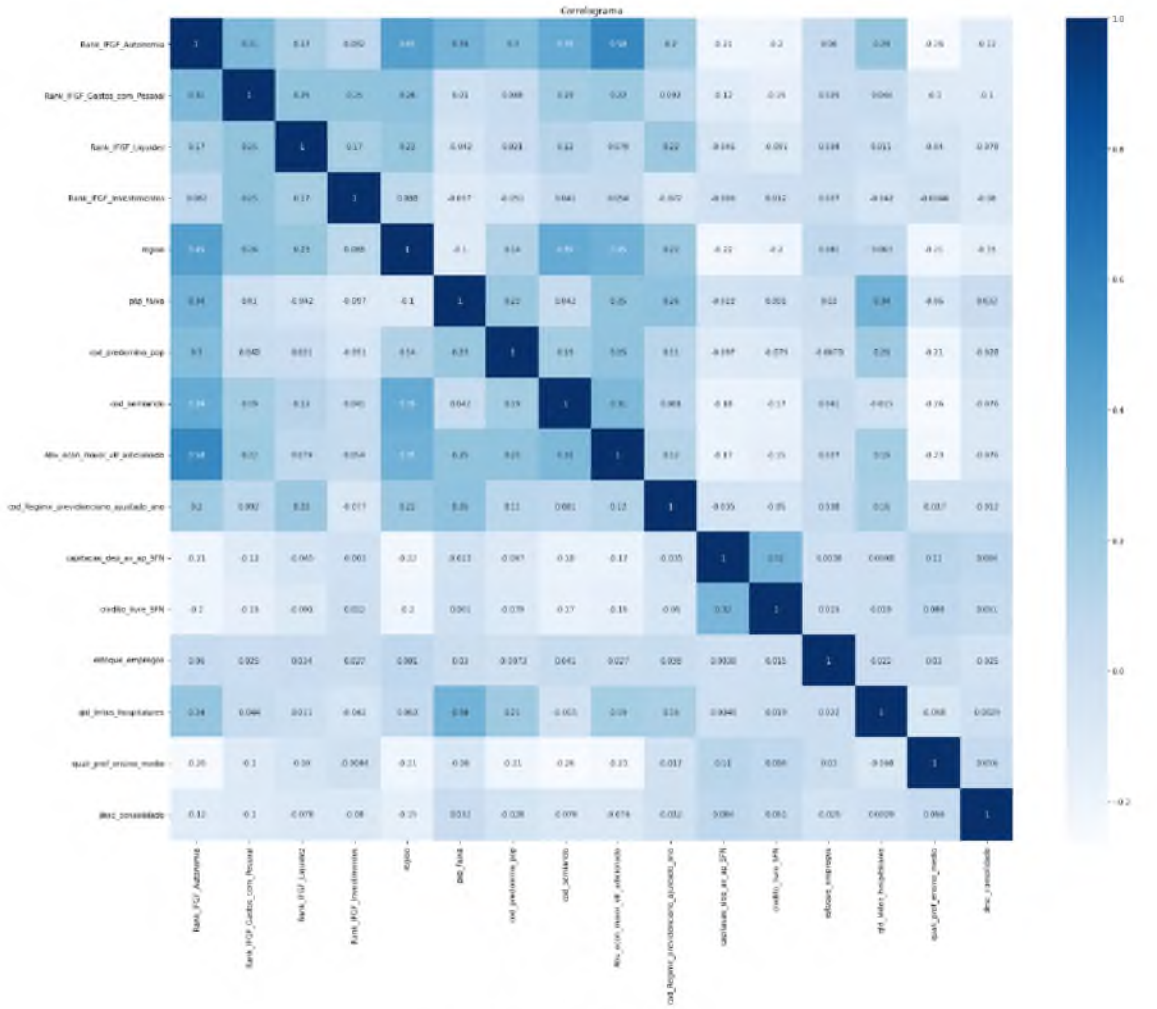
Na análise de correlação das variáveis (Figura 4.1), foram identificadas algumas variáveis com elevada correlação (acima de 60%). Após avaliação das correlações foram excluídas as seguintes variáveis por apresentarem, além de alta correlação com mais de uma variável, características semelhantes ou por apresentar pouco dinamismo, como por exemplo a variável "cod_prosp_pop". Esta variável em específico, apesar de conter informação importante sobre o nível de desenvolvimento humano e de vulnerabilidade social, não possui atualização dentro dos anos da amostra de dados, sendo a informação mais recente a disponibilizada no censo de 2010.



(Fonte: Autoria própria)

Figura 4.1: *Correlograma das variáveis*

Após o procedimento de retirada das variáveis: "cod_prosp_pop", "diversidade_econômica", "qtd_empresas", "qtd_veículos_leves", "qtd_veículos_pesados", "conexão_banda_larga" o Correlograma ficou com a seguinte conformação (Figura 4.2).



(Fonte: Autoria própria)

Figura 4.2: *Correlograma após exclusão de variáveis com alta correlação*

Considerando que a proposta de modelagem parte de um método supervisionado, em que há a informação das classes a serem preditas (*target*), optou-se por mensurar a importância das variáveis, para avaliar como está a distribuição dos atributos frente à *target* antes da aplicação dos algoritmos, no sentido de verificar se há concentração de importância em uma ou poucas variáveis. Na literatura há várias formas de mensurar o nível de importância das variáveis, segundo Archer e Kimes [96] métodos, como a análise discriminante linear, requerem que o espaço do preditor seja substancialmente reduzido antes de derivar o classificador. O algoritmo de *Random Forest* (RF) [71], não requer a redução do espaço do preditor antes da classificação. Além disso, RF produz medidas de importância variável para cada preditor candidato. Segundo os pesquisadores, foram examinadas a eficácia das medidas de importância variável do método RF na identificação do verdadeiro preditor, entre um grande número de preditores candidatos. Concluindo que o RF é adequado para uso em problemas de

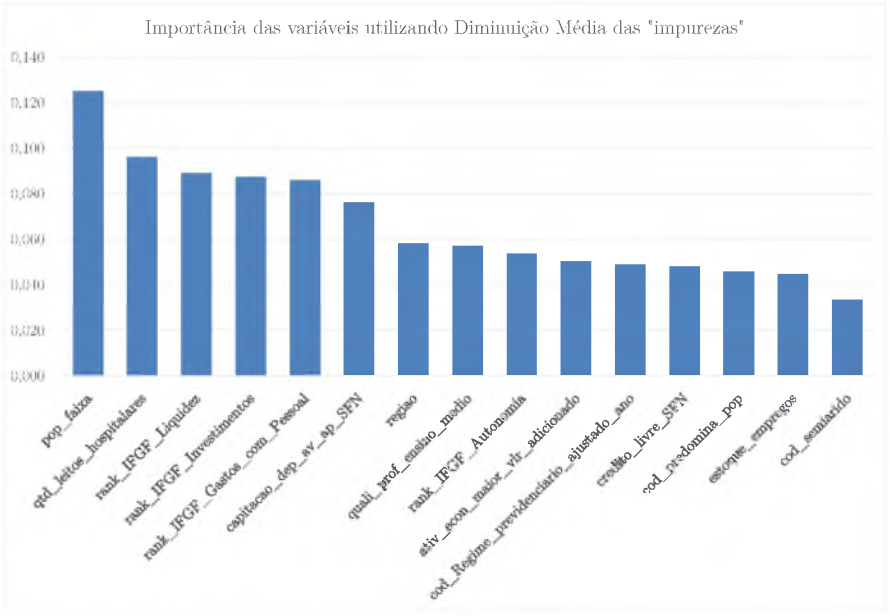
classificação quando os objetivos do estudo são produzir um classificador preciso e fornecer uma visão sobre a capacidade discriminativa de variáveis preditoras individuais.

Considerando esse referencial, foi realizada a avaliação do ganho de informação das variáveis utilizadas a partir do algoritmo *Random Forest*, com base na biblioteca *Python scikit-learn*[19]. As opções disponíveis na biblioteca são: Importância da variável, com base na diminuição média da "impureza"; e, Importância da variável, com base na permutação das variáveis. Foram realizados os testes de nas duas opções, cujos resultados estão listados a seguir:

a) Importância da variável, com base na diminuição média da "impureza" (Gráfico 4.1): Neste método as variáveis numéricas apresentaram grau maior de importância do que as variáveis categóricas. Esse fato se dá porque a importância das variáveis baseadas em "impurezas[20]", podem levar a uma superestimação da relevância daquelas que possuem alta cardinalidade (normalmente variáveis numéricas) em vez de variáveis de baixa cardinalidade, como variáveis binárias ou variáveis categóricas com um pequeno número de categorias possíveis. No contexto da base de dados, observa-se uma distribuição com pouca concentração de importância, a variável que se destaca é a "pop_faixa", relacionada às classes de população (categorias). Com o processo de transformação das variáveis categóricas em "*dummy*" e a aplicação dos algoritmos na modelagem, os valores de importância das variáveis podem ser alterados.

[19] https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html?highlight=information%20value

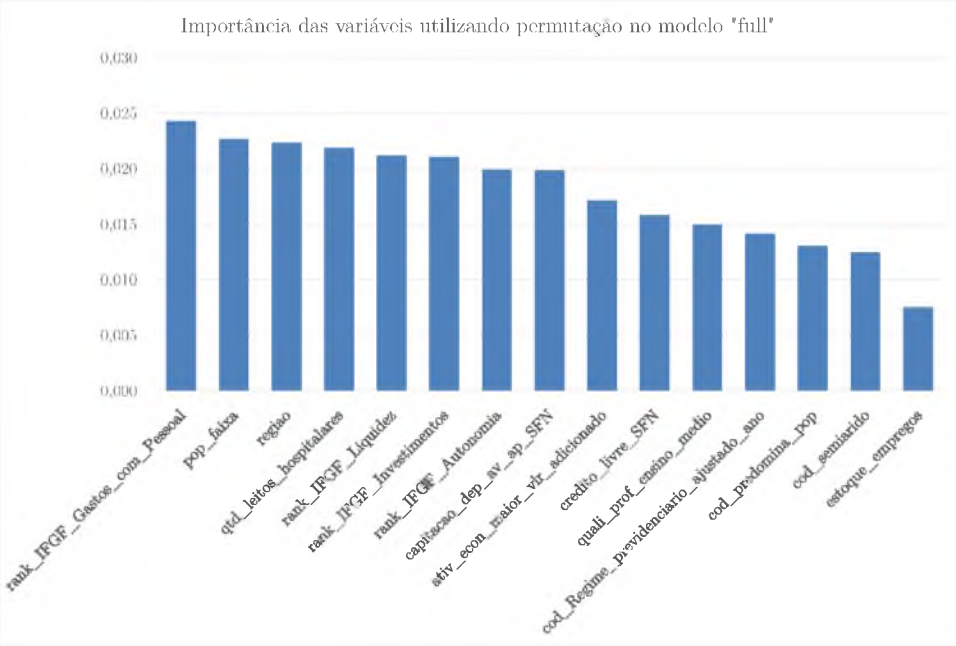
[20] A "impureza" é quantificada pelo critério de divisão das árvores de decisão (Gini, Entropia ou Erro Quadrático Médio).



Fonte: Autoria própria

Gráfico 4.1: *Importância das variáveis, utilizando Random Forest e a Diminuição Média das Impurezas*

b) Importância da variável, com base na permutação das variáveis (Gráfico 4.2): A importância da variável de permutação é definida como a diminuição na pontuação de um modelo, quando um único valor de recurso é embaralhado aleatoriamente. Tal procedimento quebra o relacionamento entre a variável e a *target*. Nesse sentido, a redução da pontuação do modelo é um indicativo do quanto o modelo depende da variável. O cálculo da importância da variável na permutação total ("*full*") é mais onerosa. As variáveis são embaralhadas "*n*" vezes e o modelo é reajustado para estimar sua importância. Assim os tipos de variáveis (categóricas ou numéricas) não recebem maior importância na definição da importância. No contexto da base de dados, observa-se também uma distribuição com pouca concentração, mas com alteração do nível de importância entre as variáveis. Como citado no item anterior, com o processo de transformação das variáveis categóricas em "*dummy*" e a aplicação dos algoritmos na modelagem, os valores de importância das variáveis podem ser alterados.



Fonte: Autoria própria

Gráfico 4.2: *Importância das variáveis, utilizando Random Forest e o método de permutação no modelo*

Para efeito do estudo, não foram excluídas as variáveis com menor grau de importância.

4.1.2. Descaracterização dos dados

A instituição financeira realizou procedimento de descaracterização dos dados originais relacionados ao Quadro 8.5 do Apêndice, além de não disponibilizar maiores detalhes sobre os critérios utilizados na caracterização dos municípios "bons" e "ruins", para a definição do *default*. Com relação aos demais dados abertos, não houve necessidade de realizar ajustes uma vez que as informações utilizadas são públicas.

Quanto à qualidade dos dados disponibilizados pela IF e das bases utilizadas para construção da base de modelagem, não foram identificadas inconsistências nos dados que levassem ao comprometimento dos resultados do estudo. Isso se deveu ao trabalho inicial de análise das bases de dados abertas e também nas avaliações das variáveis e tratamento de dados "*missing*".

4.1.3. Variável *target* de risco de crédito

A definição da variável *target*, que caracteriza o risco de crédito, foi construída a partir do conceito contido na definição de *default* do padrão internacional do Comitê de Basileia [4][5], em que na versão de Basileia II foram apresentados os requisitos para desenvolvimento dos modelos internos de risco de crédito. O referencial normativo mais atual no Brasil sobre os requisitos de gestão de risco de crédito está contido na Resolução

CMN nº 4.557/17 [6], também conhecida como GIR, Gestão Integrada do Risco e de Capital. Para efeito deste estudo, serão utilizados os requisitos contidos nessa norma, para orientar a definição da variável explicada (*target*), conforme a seguir:

> *"Art. 24. Para fins do gerenciamento do risco de crédito, a exposição deve ser caracterizada como ativo problemático quando verificado pelo menos um dos seguintes eventos:*
>
> *I - a respectiva obrigação está em atraso há mais de noventa dias;*
>
> *II - há indicativos de que a respectiva obrigação não será integralmente honrada sem que seja necessário recurso a garantias ou a colaterais.*
>
> *§ 1º Os indicativos de que uma obrigação não será integralmente honrada incluem:*
>
> *I - a instituição considera que a contraparte não tem mais capacidade financeira para honrar a obrigação nas condições pactuadas;*
>
> *II - a instituição, independentemente de exigência regulamentar, reconhece contabilmente deterioração significativa da qualidade do crédito do tomador ou contraparte;*
>
> *III - a operação relativa à exposição é reestruturada, nos termos do art. 21, § 1º, inciso II;*
>
> *IV - a instituição pede a falência ou toma providência similar em relação à contraparte; e*
>
> *V - a contraparte solicita ou sofre qualquer tipo de medida judicial que limite, atrase ou impeça o cumprimento de suas obrigações nas condições pactuadas.*
>
> *§ 2º As exposições caracterizadas como ativos problemáticos somente podem ter essa condição alterada diante de evidência de retomada, pela contraparte, da capacidade de honrar suas obrigações nas condições pactuadas. (...)"*

A identificação dos municípios "bons" ou "ruins" para efeito deste estudo teve como referência informações coletadas em bases de dados abertas e bases disponibilizadas por instituição financeira de grande porte. Esses eventos são situações que se enquadram no conceito da Resolução CMN nº 4.557/17, sinalizando a dificuldade do ente público em honrar com as responsabilidades assumidas (municípios "ruins"), tais como: atrasos em pagamentos dos compromissos financeiros, honra de garantia prestada pela União, ações judiciais movidas por instituições financeiras etc.

A prospecção de eventos que se enquadrassem no conceito e "ruins" relacionados aos municípios se baseou em informações de dados abertos ou disponibilizados pela IF de grande porte:

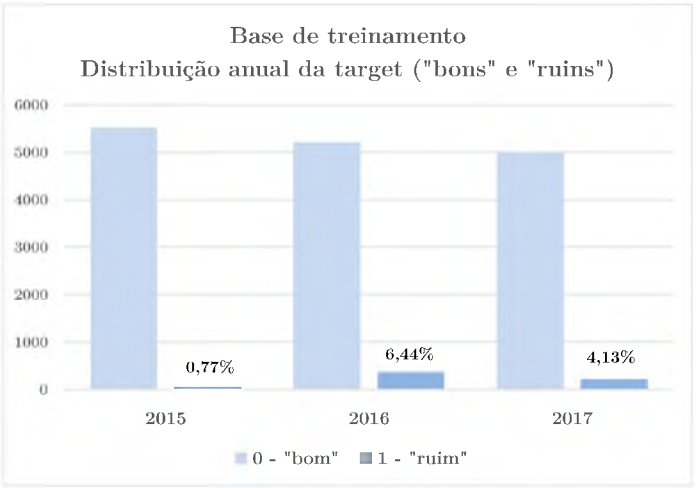
a) Tesouro Nacional – Relatório de garantias honradas pela união em operações de crédito[21];

b) Atraso de mais de 90 dias das operações de crédito na IF de referência;

c) Evento qualitativo – critério da IF[22];

d) Operações em prejuízo na IF referente ao primeiro ano de acompanhamento dos municípios.

Na construção das bases de treinamento e teste, foram observadas ocorrências de *default* que se repetiam pelo mesmo motivo, nas amostras dos quatro anos. De modo a não gerar "dupla contagem" de *default*s e também para evitar o viés no treinamento dos modelos, foi considerado na construção da base um único evento de *default*, referente ao mesmo motivo, sendo que a primeira ocorrência seria mantida na base, sendo excluídos as observações dos anos seguintes, caso fossem pelo mesmo evento. Ao final a base de treinamento ficou com 16.351 registros, sendo que a proporção entre as classes de "*default*" e "não *default*" ("bons" e "ruins"), nas visões de cada ano e também na base acumulada (empilhada), ficaram assim distribuídas conforme o Gráfico 4.3 e Gráfico 4.4 a seguir.

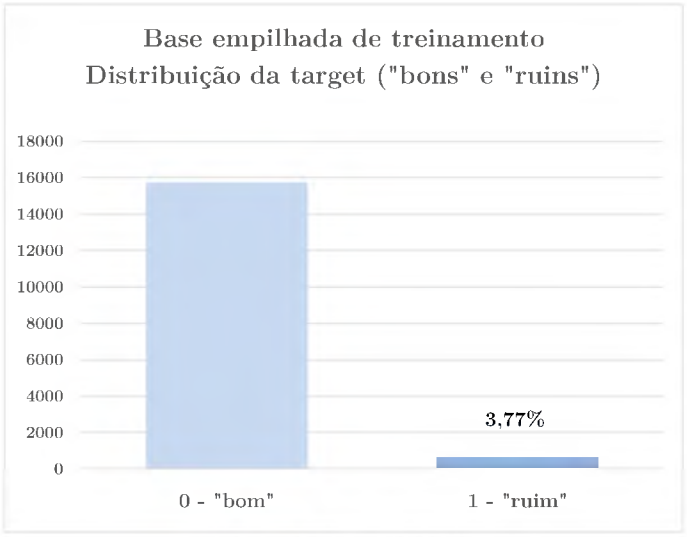


Fonte: Autoria própria

Gráfico 4.3: *Distribuição dos municípios por ano na target – "bons" e "ruins"*

[21] https://sisweb.tesouro.gov.br/apex/f?p=2501:9::::9:P9_ID_PUBLICACAO:31719

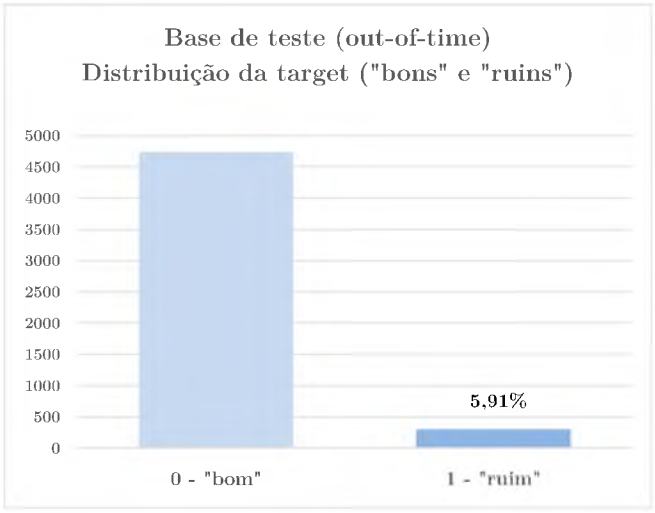
[22] Eventos observados em bases de dados especializadas de *bureaus* de crédito ou de cadastros de inadimplência.



Fonte: Autoria própria

Gráfico 4.4: *Distribuição acumulada "empilhada" dos municípios na target – "bons" e "ruins"*

A base de dados utilizada para teste dos modelos é a de 2018 (*out-of-time*), preparada com 5.029 registros (municípios), considerando os *default*s observados ao longo de 2019, que não guardam relação direta com os *default*s observados nos anos anteriores, sendo excluídos os eventos repetidos. A base de 2018 apresenta a seguinte distribuição de dados, disposta no Gráfico 4.5.



Fonte: Autoria própria

Gráfico 4.5: *Distribuição da base "out-of-time" dos municípios na target – "bons" e "ruins"*

As bases de modelagem de risco de crédito são, por natureza, desbalanceadas. Observamos tanto na base de treino como na de teste a característica de "*low default*" portfólio, dada a baixa frequência de indivíduos na classe "1 – ruim". Para tratamento

desse forte desbalanceamento, foram utilizadas técnicas para possibilitar o desenvolvimento de modelos que oferecem uma melhor predição de eventos de *default*.

4.1.1. Descrição das variáveis utilizadas na modelagem

O Quadro 4.1 apresenta as variáveis da base de modelagem utilizadas no processo de desenvolvimento e teste dos modelos para mensuração do risco de crédito dos municípios, após o pré-processamento descrito anteriormente.

Quadro 4.1: *Descrição das Variáveis*

#	Variáveis	Categoria	Tipo	Descrição
1	rank_IFGF_Autonomia	Categórica	inteira	Evidencia um dos pontos mais críticos para a gestão fiscal eficiente das prefeituras: a baixa capacidade de se sustentarem. Obs.: 5 categorias, sendo uma delas para valores 'missing'
2	rank IFGF Gastos com Pessoal	Categórica	inteira	A despesa com pessoal é o principal item da despesa do setor público, no caso dos municípios representam metade da Receita Corrente Líquida (RCL), em média. Obs.: 5 categorias, sendo uma delas para valores 'missing'
3	rank_IFGF_Liquidez	Categórica	inteira	Verifica se os recursos financeiros são suficientes para fazer frente às despesas que foram postergadas para o ano seguinte. Obs.: 5 categorias, sendo uma delas para valores 'missing'
4	rank IFGF Investimentos	Categórica	inteira	Mede a parcela dos investimentos nos orçamentos municipais. Obs.: 5 categorias, sendo uma delas para valores 'missing'
5	regiao	Categórica	inteira	macrorregiões brasileiras. Obs.: 5 categorias
6	cod_predomina_pop	Categórica	inteira	Informações por situação de domicílio. Obs.: '1 - rural' ou '2 - urbana'
7	cod_semiarido	Categórica	inteira	Informação se o município está na região semiárida. Obs.: '1 - sim' ou '2 - não'
8	ativ_econ_maior_vlr_adicionado	Categórica	inteira	Atividade econômica com maior valor adicionado bruto. Obs.: 10 categorias
9	pop_faixa	Categórica	inteira	população censo 2010. Obs.: 8 categorias
10	cod Regime previdenciario ajustado ano	Categórica	inteira	Regime Previdenciário. Obs.: '1 - RPPS', '2 - RGPS', '3 - RPPS em extinção'
11	capitacao_dep_av_ap_SFN	Numérica	inteira	Captação (depósito a prazo + poupança) no SFN
12	credito_livre_SFN	Numérica	inteira	Crédito livre no SFN
13	estoque_empregos	Numérica	inteira	Estoque de empregos do setor formal, ponderado pela participação dos cinco setores em cada município
14	qtd_leitos_hospitalares	Numérica	inteira	Número de leitos hospitalares (complementares + internação)
15	quali_prof_ensino_medio	Numérica	inteira	Qualificação profissional (percentual de trabalhadores formais com ensino médio completo)
16	desc_consolidado	Categórica	inteira	Variável target "não default" (0) e "default"(1)

Fonte: Autoria própria

4.2. Modelagem

Nesta etapa está descrito o "passo-a-passo" do desenvolvimento dos modelos e os seus resultados do treinamento, para se chegar no cálculo de uma probabilidade de relacionada ao risco de crédito dos municípios. Lembrando que para se estimar a probabilidade foram aplicadas três técnicas distintas: (i) *XGBoost*, (ii) *Random Forest* e (iii) Regressão Logística, em três alternativas de modelagem, sendo duas relacionadas às melhores práticas para situações de desbalanceamento de classes.

4.2.1. Tratamento das variáveis para aplicação dos algoritmos

No processo de modelagem utilizando algoritmos de *machine learning* faz-se necessário ajustar as variáveis em uma formatação estrutura de informação que seja "legível" para os algoritmos, de modo a não considerar um valor categórico mais importante do que outro. Por exemplo, na variável "regiao" há os seguintes valores categorizados: "1", "2", "3", "4" e "5", representando as cinco regiões do Brasil, para que os algoritmos não considerem o valor "5" mais importante do que os demais, utiliza-se um processo de transformação dos valores das variáveis categóricas. Um ajuste possível é transformar as variáveis categóricas em variáveis "*dummy*". Vale destacar que as variáveis categóricas do IFGF não passaram por esse tratamento, para que a gradação dos indicadores (5–"Gestão de Excelência", 4–"Boa Gestão", 3–"Gestão em Dificuldade", 2–"Gestão Crítica", 1–"Não possui IFGF") fosse capturada pelos modelos.

No processo de preparação das bases de treinamento e teste, foi aplicado o pacote *sklean OneHotEncoder*[23] para realizar a transformação das variáveis.

Quadro 4.2: *transformação das variáveis categóricas em dummy*

```
...
# exemplo de registro (0) da base de treinamentc antes do processo transformação em
dummy
X[0]
array([4, 5, 4, 2, 5, 2, 2, 5, 1, 2, 2, 2, 2, 3, 2])

# exemplo de registro (0) da base de treinamentc após o processc transformação em
dummy
X[0]
array([0., 0., 0., 0., 1., 0., 1., 0., 1., 0., 0., 0., 0., 1., 0., 0., 0.,
       0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 1., 0., 0.,
       1., 0., 1., 0., 0., 0., 1., 0., 0., 1., 0., 4., 5., 4., 2.])
...
```

Fonte: Autoria própria

[23] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

O próximo passo de ajuste das variáveis é padronizar os valores numa escala. Para realizar a transformação foi utilizada o pacote *sklearn* de pré-processamento *StandardScaler*[24] na base de treinamento e teste.

Quadro 4.3: *Padronização das variáveis em escala*

```
...
# exemplo de registro (0) da base de treinamento após o processo padronização
X[0]
array([-0.29204689, -0.67893612, -0.66044528, -0.52775051,  3.29501788,
       -0.63864686,  0.63864686, -0.50026755,  0.50026755, -0.9881444 ,
       -0.36292387, -0.09622208, -0.04141704,  1.65501683, -0.1186524 ,
       -0.23212785, -0.10986273, -0.17446667, -0.06923305,  1.12235193,
       -0.57172411, -0.34763883, -0.24417257, -0.17576515, -0.25820731,
       -0.17032897, -0.16453834, -1.25902277,  1.26881687, -0.06068785,
       -1.19233084,  1.61984315, -0.39842365, -1.0179583 ,  1.0179583 ,
       -0.30312782,  0.3224543 , -0.10065579, -0.74330791, -0.34597895,
        1.54421484, -0.56466074, -0.17465269,  0.56390717, -0.51817248,
        0.91259509,  1.76233483,  0.71001546, -0.72325573])
...
```

Fonte: Autoria própria

Essa etapa consiste em utilizar a distribuição normal e seus parâmetros (média zero e variância constante) para transformar os valores das variáveis da base de dados de modelagem, convertendo essas observações em um intervalo específico, podendo ser [-1,1] ou [0,1]. A padronização de um conjunto de dados é um requisito comum para muitos estimadores de aprendizado de máquina, dado que esses algoritmos podem se comportar mal se as variáveis não tiverem um comportamento semelhante à distribuição normal.

Por exemplo, muitos elementos usados na função objetivo de um algoritmo de aprendizagem (como o *kernel* RBF de *Support Vector Machines* ou os regularizadores L1 e L2 de modelos lineares) assumem que todos os recursos estão centrados em torno de 0 e têm variância na mesma ordem. Se uma variável possui variância com magnitude maior do que as outras, esta pode dominar a função objetivo e tornar o estimador incapaz de aprender com outras variáveis corretamente como esperado, "viesando" o modelo.

No processo de modelagem a base de dados utilizada para treinamento (2015, 2016 e 2017) foi adotado o procedimento de *cross-validation*, em que a base de dados foi estratificada com três repetições de 10 "estratos" (*fold*), utilizado a técnica de *k-fold*, gerando 30 amostras estratificadas para treinamento dos modelos. Para avaliar os modelos, foi utilizada a curva ROC AUC, sendo calculada a pontuação média das amostras.

Conforme H.He e Y.Ma [97], há que se ter atenção no uso do *cross-validation* (10-fold), particularmente quando os dados estão com forte desbalanceamento de classes, pois

[24] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

esse método pode facilmente ser "quebrado". Nesse caso, a solução é não dividir os dados aleatoriamente ao usar o *cross-validation k-fold* ou uma divisão de treino e teste.

Para minimizar o problema, Brownlee [59] propõe processo em que a divisão aleatória do conjunto de dados aconteça de forma a manter a mesma distribuição de classes em cada subconjunto. A estratificação ou amostragem estratificada é realizada tendo a variável *target* (y) como controladora do processo de amostragem. Dessa forma, podemos usar uma versão de *cross-validation k-fold* que preserva a distribuição de classe desequilibrada em cada estrato (*fold*). A biblioteca utilizada para realizar estratificação e repetição da base de treinamento e teste foi a *sklearn.RepeatedStratifiedKFold*[25].

Os algoritmos escolhidos para classificar os municípios "bons" e "ruins", com informação de probabilidade, considerando o referencial teórico foram o *XGBoost*[26], o Randon Forest[27] e a Regressão Logística[28], a partir da API *XGBoost* e das bibliotecas disponíveis da scikit-learn.org, em *Python*.

Conforme observado por Brownlee [59], poucos algoritmos de aprendizagem de máquinas produzem probabilidades calibradas, como por exemplo: Regressão Logística, Análise Discriminante Linear, *Naive Bayes*, Redes Neurais. Essa situação acontece porque para um modelo prever probabilidades calibradas, ele deve ser explicitamente treinado em um contexto probabilístico, como a máxima estimativa de verossimilhança.

Muitos outros algoritmos, como SMV, Árvores de decisão, *Bagging*, *Random Forest*, Gradiente *Boosting*, KNN estimam um score semelhante a uma probabilidade ou um "rótulo" para classe. Para utilizá-los, faz-se necessário que as pontuações geradas pelos modelos, com base nesses algoritmos, sejam calibradas para uma probabilidade antes de serem aplicados [98]. Assim as probabilidades dos modelos propostos foram calibradas, além de reescalar seus valores, para que melhor combinem com a distribuição observada nos dados de treinamento. Para realizar a calibragem dos modelos desenvolvidos com base no *Random Forest* e *XGBoost*, foi adotada a biblioteca *Sklearn.CalibratedClassifierCV*[29], com parametrização do método "*sigmóide*", que corresponde ao método de Platt (ou seja, um modelo de regressão logística).

---

[25] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RepeatedStratifiedKFold.html
[26] https://xgboost.readthedocs.io/en/latest/python/python_api.html
[27] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[28] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[29] https://scikit-learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html

4.2.2. Modelos desenvolvidos a partir de dados desbalanceados

Nesta etapa foram aplicados os algoritmos *XGBoost*, *Random Forest* e Regressão Logística diretamente nas bases desbalanceadas, sem qualquer ajuste de parâmetros ou hiperparâmetros.

Os resultados dos indicadores de performance para os modelos desenvolvidos estão apresentados na Tabela 4.3:

Tabela 4.3: *Desempenho dos modelos no treinamento – Base desbalanceada*

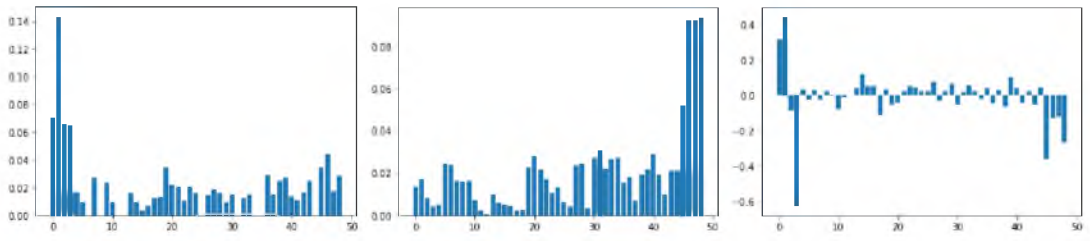
	XGB	RF	RL
AUC	0,762	0,989	0,736
Acurácia	0,716	0,984	0,700
F1	0,178	0,827	0,163
Precisão	0,100	0,708	0,091
Recall	0,812	0,994	0,775
KS	0,526	0,977	0,478

Legenda: XGB – *XGBoost*; RF – *Random Forest*; RL – Regressão Logística

Fonte: Autoria própria

Observamos aqui valores elevados dos indicadores de desempenho no modelo treinado com base no algoritmo *Random Forest*. Os demais modelos não apresentaram valores destacados.

As importâncias das variáveis (Figura 4.3) tiveram comportamentos distintos nos modelos: no caso do modelo XGB, o conjunto de opções do atributo "região" teve um peso maior no resultado do modelo; já no RF as variáveis de IFGF foram as mais importantes. No RL, as variáveis que tiveram maior importância foram a "região", seguida das variáveis IFGF.



XGB RF RL

Fonte: Autoria própria

Figura 4.3: *Importância das variáveis dos Modelos Base Desbalanceada*

No caso da RL, os coeficientes (β) relacionados às variáveis são positivos e negativos. As pontuações positivas indicam uma característica que prevê a classe 1, enquanto as pontuações negativas indicam uma característica que prevê a classe 0. Como as variáveis estão normalizadas/padronizadas, os β trazem um referencial de importância das variáveis na Regressão Logística.

4.2.3. Modelos desenvolvidos considerando a sensibilidade ao custo

A maioria dos algoritmos classificadores assumem que os custos de uma classificação incorreta (falso negativo e falso positivo) possuem o mesmo peso ou custo. No entanto, na maioria das aplicações no mundo real, essa suposição não é verdadeira [63]. Particularmente, no caso deste estudo não foi possível capturar informações sobre o custo de um município ser classificado como "bom" ou "ruim" para ponderar na modelagem. Dadas as limitações no uso de outras formas de medir o custo da classificação incorreta, a análise de sensibilidade ao custo se restringiu à metodologia de "matriz de custo", em que se apurou uma relação de custo a partir da matriz de confusão. O modo "balanceado" utiliza os valores de "y" (0,1), para ajustar automaticamente os pesos inversamente proporcionais às frequências de classe nos dados de entrada.

Na aplicação dos algoritmos nessa etapa, adotou-se o custo "balanceado" entre as classes na parametrização. Resultados dos indicadores de performance para os modelos desenvolvidos estão dispostos na Tabela 4.4:

Tabela 4.4: *Desempenho dos modelos no treinamento – Sensível ao custo*

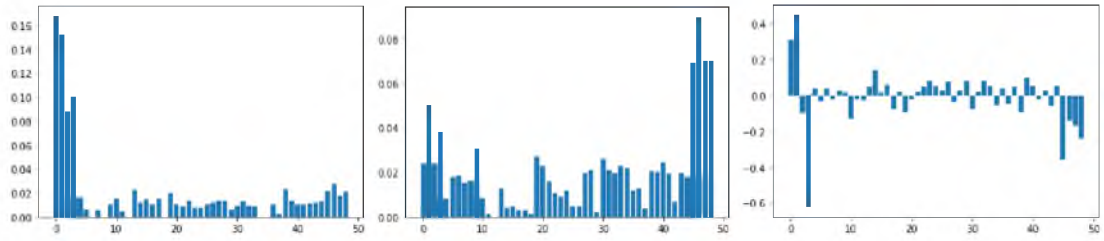
	XGBsc	RFsc	RLsc
AUC	0,770	0,994	0,742
Acurácia	0,738	0,988	0,678
F1	0,188	0,867	0,160
Precisão	0,107	0,766	0,089
Recall	0,804	1,000	0,810
KS	0,546	0,988	0,485

Legenda: XGB – *XGBoost*; RF – *Random Forest*; RL – Regressão Logística

Fonte: Autoria própria

Novamente, observamos valores elevados no modelo treinado com base no algoritmo *Random Forest*, caracterizando possibilidade de "sobre ajuste" (*overfitting*). Os demais modelos não apresentaram valores discrepantes.

As importâncias das variáveis (Figura 4.4), assim como nos modelos anteriores, tiveram comportamentos distintos nos modelos, no caso do modelo XGB o conjunto de opções do atributo "região" teve um peso maior no resultado do modelo, já no RF as variáveis de IFGF foram as mais importantes. No RL, as variáveis que tiveram maior importância foram a "região", seguida das variáveis IFGF.



XGBsc RFsc RLsc

Fonte: Autoria própria

Figura 4.4: *Importância das variáveis dos Modelos Sensíveis ao Custo*

Adicionalmente, foi realizado um "*GridSearch*[30]" para avaliar a sensibilidade do modelo à relação de peso entre as classes, buscando identificar qual parâmetro de custo traria o melhor resultado de AUC, utilizando as seguintes proporções de classes:

Tabela 4.5: *Pesos para análise com base no custo*

Modelo	Custo parametrizado
XBGsc	[1, 10, 25, 50, 75, 99, 100, 1000, 'balanceada']
RFsc	[{0:100,1:1}, {0:10,1:1}, {0:1,1:1}, {0:1,1:10}, {0:1,1:100}, 'balanceada']
RLsc	[{0:100,1:1}, {0:10,1:1}, {0:1,1:1}, {0:1,1:10}, {0:1,1:100}, 'balanceada']

Legenda: XGBsc – *XGBoost*; RFsc – *Random Forest*; RLsc – Regressão Logística

Fonte: Autoria própria

Os resultados das "melhores" pontuações de AUC nas configurações de pesos, não tiveram diferenças relevantes frente à referência "balanceada".

[30] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Tabela 4.6: *Desempenho dos modelos – Base desbalanceada*

Peso das classes	XGBsc	**Classe:Peso**	RFsc	RLsc
1	0.7903*	0:100, 1:1	0.7178	0.7830
10	0.7896	0:10, 1:1	0.7190*	0.7830
25	0.7888	0:1, 1:1	0.7175	0.7832*
50	0.7854	0:1, 1:10	0.7128	0.7831
75	0.7833	0:1, 1:100	0.7153	0.7815
99	0.7831	'balanceada'	0.7152	0.7828
100	0.7827			
1000	0.7751			
'balanceada'	0.7879	(*) melhor resultado		

Fonte: Autoria própria

Observou-se que no caso do RFsc, o algoritmo escolheu uma relação que aumenta em 10 vezes o peso da classe "0" frente a classe "1": (0:10, 1:1), gerando um modelo que prevê muito mais a classe "0" do que a "1", algo que não é desejado no contexto de avaliação do risco de crédito.

4.2.4. Modelos desenvolvidos a partir de dados balanceados

Para minimizar o problema da base desbalanceada, na aplicação dos algoritmos *XGBoost* e *Random Forest*, foram testados quatro métodos de reamostragem a partir do referencial teórico. Cabe destacar que não existe um único método melhor para aplicar em problemas de *low default*. Como na escolha de um modelo preditivo, se faz necessário realizar uma experimentação cuidadosa, na busca de descobrir o que funciona melhor para cada problema, e é nesse contexto que foram realizadas reamostragens utilizando métodos combinados de *Oversample* e *Undersample*:

a) *RandomOverSampler*[31] e *RandomUnderSampler*[32];

b) SMOTE[33] e *RandomUnderSampler*;

c) SMOTE e *TomekLink*[34];

[31] https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html

[32] https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html

[33] https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html?highlight=smote#imblearn.over_sampling.SMOTE

d) SMOTE e *Edited Nearest Neighbors* (ENN)[35].

A parametrização dos métodos citados para geração das amostras para modelagem (treinamento) e a quantidade de observações para cada classe após a reamostragem está descrita na Tabela 4.7, a seguir:

Tabela 4.7: *Parametrização das amostras balanceadas*

	Estratégia de amostragem	Quantidade de observações
RandomOversampler e *RandomUndersampler*	over = (0.1)* under = (0.5)*	4.719
SMOTE e *RandomUndersampler*	over = (0.1)* under = (0.5)*	4.719
SMOTE e *Tomek Link*	Reamostrar a classe majoritária no *TomekLink*	31.465
SMOTE e *Edited Nearest Neighbors*	Não reamostrar a classe majoritária	28.286

Legenda: XGB – *XGBoost*; RF – *Random Forest*; RL – Regressão Logística
(*) razão entre o número de amostras das classes minoritária e a majoritária após a reamostragem.

Fonte: Autoria própria

No processo de amostragem combinada SMOTE e *TomekLink*, adotou-se estratégia semelhante à aplicada no trabalho de Batista, Bazzan e Monard [61], em que apenas exemplos de classes majoritárias que participam do *TomekLink* foram removidos, uma vez que eventos de classes minoritárias são considerados raros para serem descartados. Como os eventos de classes minoritárias foram artificialmente criados pelo método SMOTE e os conjuntos de dados estão balanceados, ambos os exemplos de classes majoritárias e minoritárias, foram removidos pelo processo *TomekLink*.

Uma vez balanceados, os algoritmos de aprendizado de máquina padrão podem ser treinados diretamente no conjunto de dados transformado sem qualquer modificação. [59]. Os algoritmos *XGBoost*, *Random Forest* e Regressão Logística foram aplicados diretamente nas bases balanceadas, sem qualquer ajuste de parâmetros ou hiperparâmetros.

[34] https://imbalanced-learn.org/stable/references/generated/imblearn.combine.SMOTETomek.html?highlight=smotetomek#imblearn.combine.SMOTETomek
[35] https://imbalanced-learn.org/stable/references/generated/imblearn.combine.SMOTEENN.html?highlight=smoteenn#imblearn.combine.SMOTEENN

Resultados dos indicadores de performance para os modelos desenvolvidos:

Tabela 4.8: *Desempenho dos modelos no treinamento – Bases balanceadas*

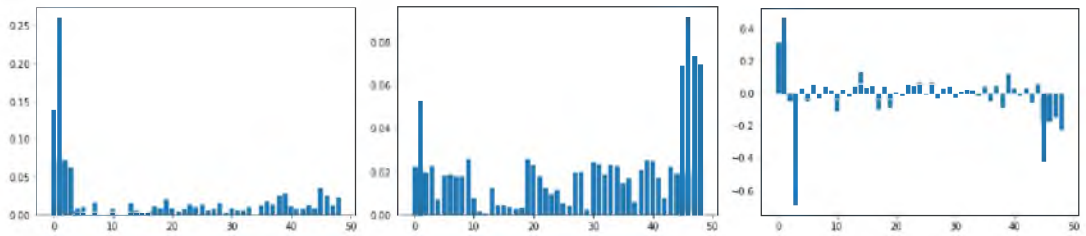
	RandomUndersample e RandomOversample			SMOTE e RandomUndersample			SMOTE e Tomek Link			SMOTE e Edited Nearest Neighbors		
	XGBou	RFou	RLou	XGBsu	RFsu	RLsu	XGBst	RFst	RLst	XGBse	RFse	RLse
AUC	0,764	0,959	0,740	0,749	0,955	0,739	0,734	0,972	0,741	0,737	0,950	0,743
Acurácia	0,707	0,944	0,655	0,717	0,948	0,654	0,689	0,972	0,654	0,715	0,958	0,646
F1	0,175	0,570	0,154	0,173	0,582	0,153	0,160	0,727	0,154	0,168	0,626	0,153
Precisão	0,098	0,402	0,085	0,097	0,417	0,084	0,089	0,581	0,085	0,094	0,469	0,084
Recall	0,825	0,976	0,831	0,783	0,963	0,830	0,783	0,972	0,835	0,762	0,942	0,848
KS	0,528	0,921	0,481	0,506	0,912	0,479	0,468	0,945	0,481	0,482	0,901	0,489

Legenda: XGB – *XGBoost*; RF – *Random Forest*; RL – Regressão Logística

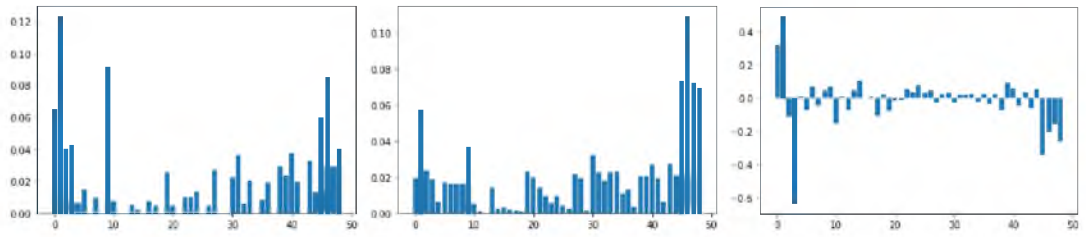
Fonte: Autoria própria

Observamos valores elevados no modelo treinado com base no algoritmo *Random Forest*, caracterizando possibilidade de sobreajuste (*overfitting*). Os demais modelos não apresentaram valores discrepantes.

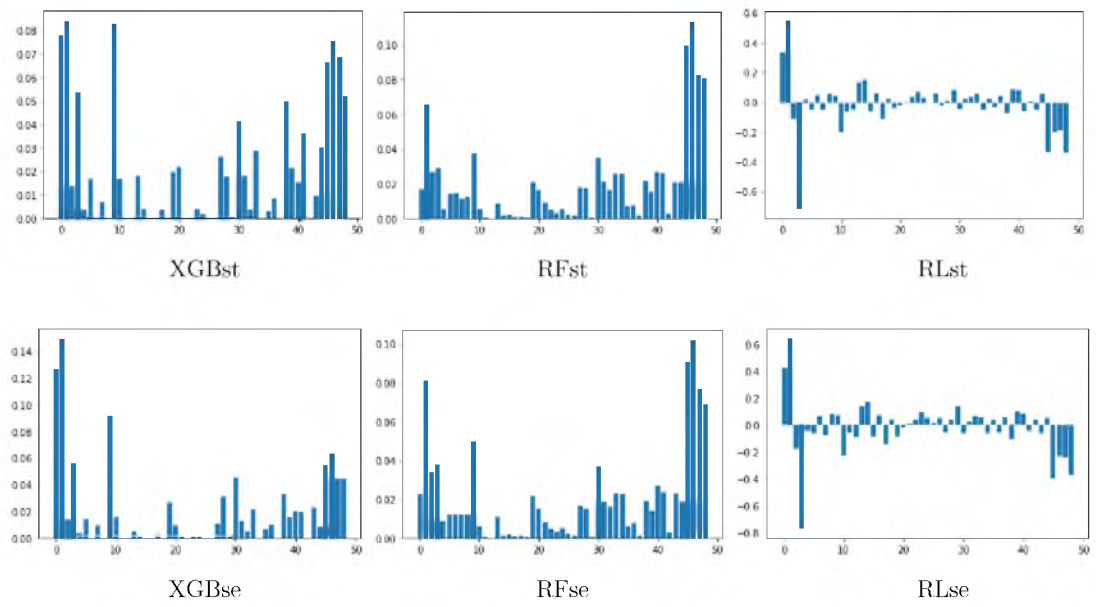
As importâncias das variáveis, destacadas na Figura 4.5, tiveram comportamentos distintos nos modelos de XGB e RF, dependendo da técnica de amostragem utilizada. Verificam-se valores diferenciados dos pesos das variáveis, mas mantendo o padrão já observado nos modelos gerados pelos métodos anteriores. No caso desses modelos, colocando o conjunto das opções do atributo "região" e as variáveis de IFGF, com uma importância maior no resultado dos modelos XGB e RF respectivamente. Há um domínio da variável "ativ_econ_maior_vlr_adicionado" se destacando em importância em três modelos XGB. No caso dos modelos RL, as variáveis em importância se mantiveram as mesmas: "região" e o conjunto IFGF, destaque para a uniformidade entre os modelos nesse quesito, variando discretamente os valores de importância.



XGBou RFou RLou



XGBsu RFsu RLsu



Fonte: Autoria própria

Figura 4.5: *Importância das variávcis dos Modelos Bases Balanceadas*

4.2.5. Resultados dos modelos desenvolvidos

Antes de apresentar as matrizes de confusão dos modelos e os resultados das curvas ROC e KS, cabe o comentário sobre os resultados das métricas de desempenho dos algoritmos aplicados na base de treinamento, destacadas em cada método citado anteriormente. Os pontos de corte utilizados para mensuração dos indicadores, exceto KS, foram os referenciados nos valores limites de probabilidade da curva ROC, considerando o valor máximo observado para o indicador *G-mean*. Segundo Brownlee [59], se a distribuição de classes for severamente desbalanceada, então a métrica *G-mean* pode ser usada para otimizar a sensibilidade e as métricas específicas. A ideia é encontrar para cada Curva ROC AUC, um valor otimizado entre a taxa de verdadeiros positivos (Sensibilidade) e taxa calculada por um menos os falsos positivos (Especificidade).

O valores de limites referenciais foram obtidos a partir das predições de classificação $\hat{y}(0,1)$ de cada um dos modelos treinados comparando com o valor de y(0,1) da base de treinamento, sem o efeito de reamostragem ou balanceamento. O resultado dos valores limites e respectivos *G-mean* estão dispostos na Tabela 4.9 a seguir.

Tabela 4.9: *Limites da curva ROC – Base de treinamento*

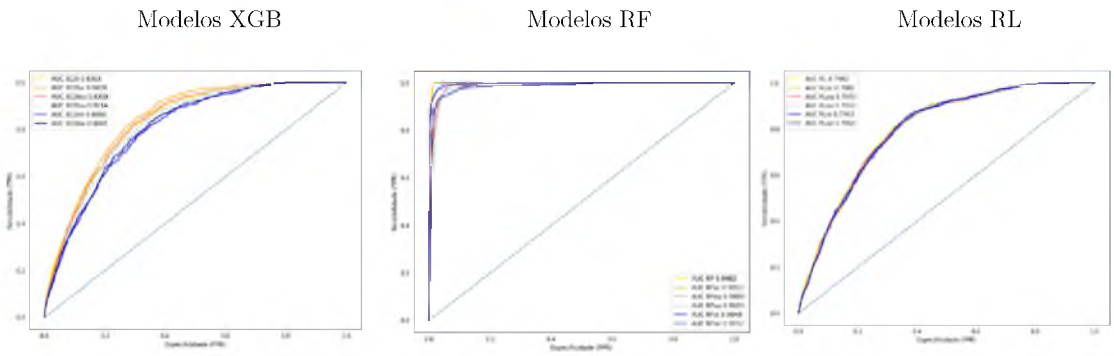
Método	Modelos	Melhor limite (probabilidade)	G-mean
Bases desbalanceadas	XGB	0.036135	0.761
	RF	0.061081	0.989
	RL	0.045181	0.735
Sensível ao custo	XGBsc	0.051010	0.770
	RFsc	0.084672	0.994
	RLsc	0.510820	0.739
Bases balanceadas			
RandomOversample e *RandomUndersample*	XGBou	0.147638	0.748
	RFou	0.384969	0.960
	RLou	0.309218	0.734
SMOTE e *RandomUndersample*	XGBsu	0.147638	0.748
	RFsu	0.405622	0.956
	RLsu	0.299373	0.734
SMOTE e *Tomek Link*	XGBst	0.023457	0.732
	RFst	0.069585	0.972
	RLst	0.447128	0.735
SMOTE e *Edited Nearest Neighbors*	XGBse	0.030273	0.737
	RFse	0.203835	0.951
	RLse	0.476433	0.736

Legenda: XGB – *XGBoost*; RF – *Random Forest*; RL – Regressão Logística

Fonte: Autoria própria

Os modelos que apresentaram o maior valor *G-mean* foram os desenvolvidos utilizando o método Sensível ao Custo. Destaque para os modelos *Random Forest* pelo desempenho bem acima dos demais classificadores, todavia há que se verificar se há comportamento de *overfitting* como comentado em outros pontos deste documento.

As curvas ROC AUC para cada conjunto de modelos desenvolvidos a partir dos algoritmos XGB, RF e RL, aplicados na base de treinamento sem tratamento apresentaram o comportamento descrito na Figura 4.6 a seguir.



Fonte: Autoria própria

Figura 4.6 *– Curvas ROC AUC dos Modelos XGB, RF e RL aplicados na base de treinamento*

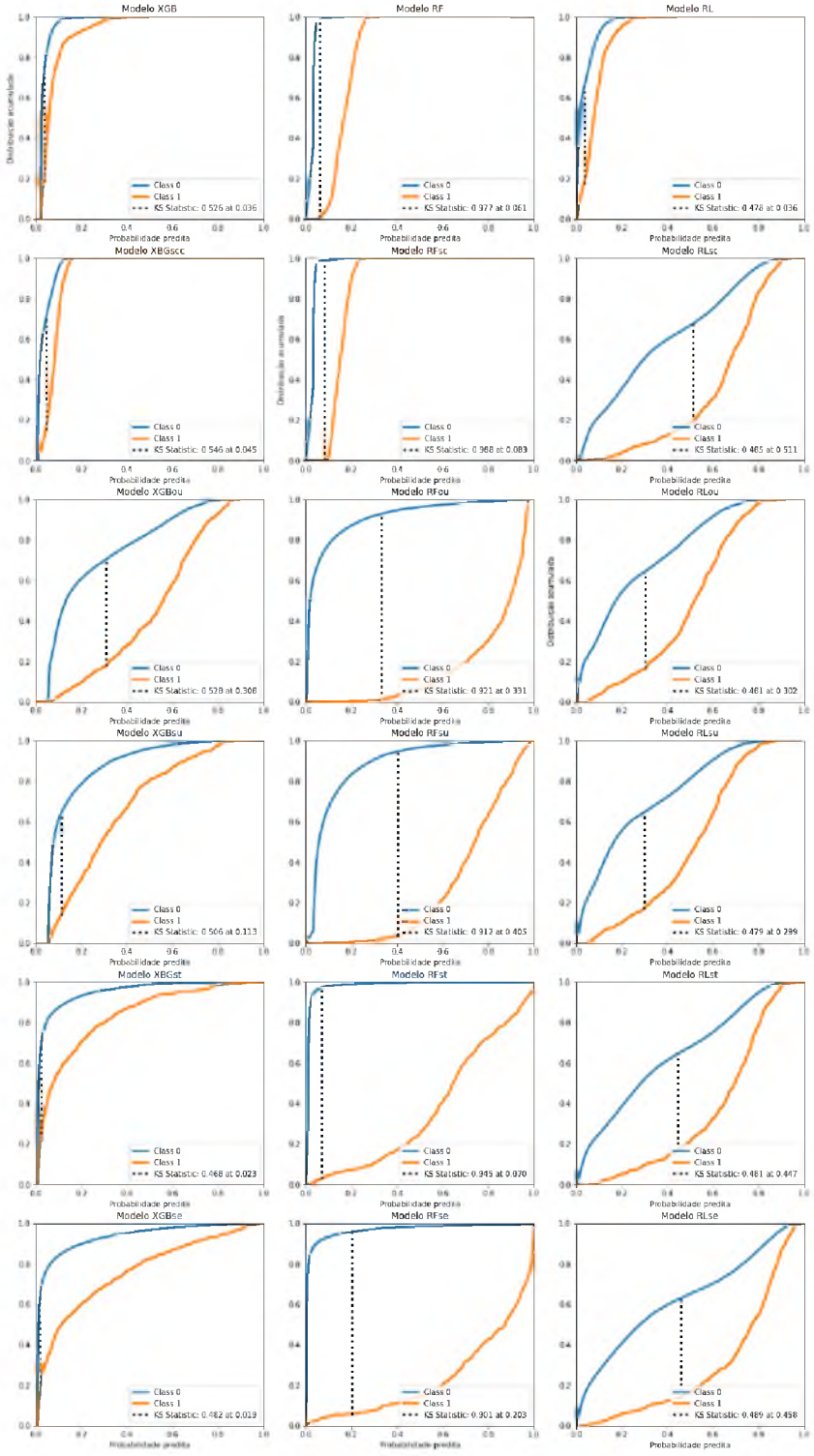
Os modelos XGB foram mais sensíveis aos métodos, alcançando uma variação maior para os valores de ROC AUC, já os modelos RF apresentaram valores elevados para todos os métodos utilizados, sinalizando que pode estar com super ajuste. Os modelos RL foram pouco sensíveis aos métodos utilizados não havendo diferenças visíveis observadas graficamente, tal fato pode estar relacionado com os mecanismos de regularização do algoritmo.

Os gráficos gerados (Figura 4.7) para verificar o comportamento das funções de distribuição acumulada das probabilidades dos "bons" e "ruins", calculadas pelo método KS (quanto maior melhor), evidenciam os resultados dos modelos com relação a discriminar os municípios classificados nas duas classes.

Assim como mostrado nas curvas ROC AUC os modelos tiveram comportamento distinto nas probabilidades calculadas para as duas classes de municípios. Os modelos RF tiveram elevados valores de KS, sendo o maior 0,988 (RFsc) e o menor 0,901 (RFse), esses valores estão muito acima dos normalmente referenciados por Sicsú [70], em que valores de KS maior do que 0,70 seriam valores "pouco usuais". Os modelos XGB tiveram na sequência os melhores valores de KS situando entre maior 0,546 (XGBsc) e o menor 0,468 (XGBst), já os modelos RL tiveram os valores de KS entre 0,478 (RL) e 0,489 (RLse) e apresentaram distribuições de probabilidade acumuladas semelhantes, como destacado nos gráficos, exceto o modelo RL gerado a partir das bases desbalanceadas.

Os pontos de maior KS traz a probabilidade de "ponto de corte" semelhante ao limite da ROC AUC, e no caso da RL seria um valor próximo à taxa de *default* encontrada na amostra utilizada no processo de treinamento. Por exemplo, a taxa de *default* da amostra desbalanceada da base de treinamento é de 3,77% de municípios "ruins" e a probabilidade estimada pelo modelo RL treinado nessa base, para o ponto de máximo KS é de 3,60%. Isso leva a crer que a diferença dos valores das probabilidades

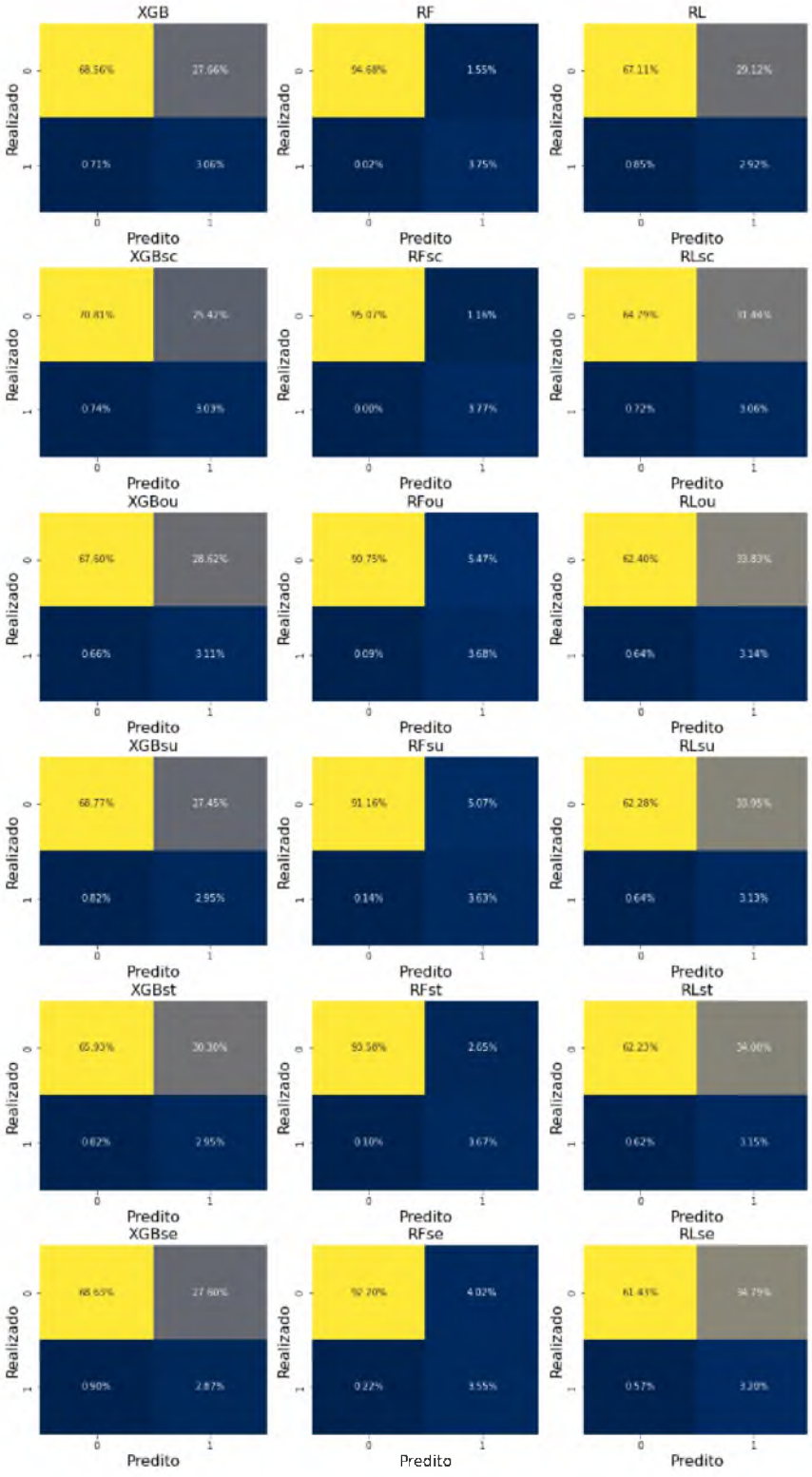
calculadas dos modelos seja explicada pelo valor distinto do intercepto da função gerada para cada modelo. Outro ponto observado, é que os coeficientes de cada uma das variáveis não variam muito entre os modelos RL, mesmo sendo treinados em amostras com características diferentes.



Fonte: Autoria própria

Figura 4.7 – *Curvas KS dos Modelos XGB, RF e RL aplicados na base de treinamento*

As matrizes de confusão (Figura 4.8) geradas a partir do ponto de corte otimizado (*G-mean*) apresentaram os seguintes comportamentos de percentuais de taxas de acerto e erros tipo I e II para cada modelo.



Fonte: Autoria própria

Figura 4.8 – *Matrizes de confusão dos Modelos XGB, RF e RL aplicados na base de treinamento*

Excetuando os modelos RF, os modelos XGB e RL tiveram baixos percentuais de erro Tipo I (falso positivo), todavia os valores de erro Tipo II (falso negativo) ficaram proporcionalmente elevados, colocando "bons" municípios como "ruins". Nesse aspecto, os modelos valorizam uma classificação mais precisa dos municípios "ruins" em detrimento daqueles entes considerados "bons".

5. Validação e Análise dos Resultados

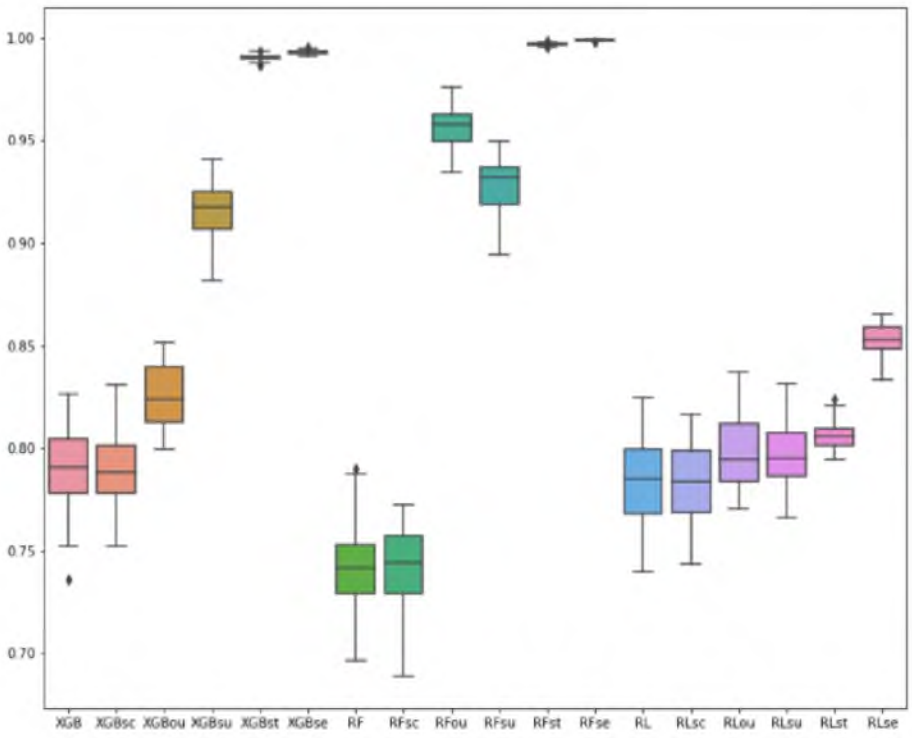
Nesta fase foram realizados testes, a partir dos resultados do *cross-validation* do ROC AUC (30 amostras de cada modelo), para verificar se os modelos gerados são diferentes entre si e quais seria os "melhores" a partir do treinamento.

Na sequência os modelos treinados foram aplicados na base de teste de 2018 (out-of-time) para avaliar a performance de cada um em dados desconhecidos. Para efeito deste estudo, o cálculo dos indicadores de desempenho dos modelos teve como referência os limites de ROC AUC calculados no desenvolvimento dos modelos. Caso seja de interesse da instituição financeira implementar os modelos, pode-se recalcular os valores dos indicadores e ponto de corte a partir do observado na base de teste.

Este capítulo se encerra com a aplicação dos "melhores" modelos a partir da avaliação geral dos resultados de treinamento e teste. Nessa etapa, foram propostas classificações em níveis de riscos a partir das probabilidades calculadas para cada modelo aplicado na base de teste, posteriormente os resultados dos modelos aplicados na base de teste foram comparados com um referencial (*benchmark*) público: CAPAG.

5.1. Teste comparativo dos modelos treinados

A partir dos resultados das amostras geradas no *cross-validation* dos modelos, tendo como referência os valores de AUC para cada amostra (30), foi gerado inicialmente um gráfico *box-plot*. A Figura 5.1 resume os resultados, sendo possível perceber que os modelos com maiores valores de AUC e menor dispersão, em tese, seriam os modelos melhores e mais estáveis, por apresentar comportamento mais uniforme independentemente da amostra de dados utilizada no treinamento. Assim podemos observar que os modelos XGBse, XGBst, RFse e RFst apresentaram melhores referenciais. No entanto, o conjunto dos modelos de Regressão Logística tiveram medianas com menor variabilidade, evidenciando que independentemente do tipo de amostra o algoritmo tem melhor adaptação. Na sequência avaliamos se os modelos possuem diferenças significativas entre si.



Fonte: Autoria própria

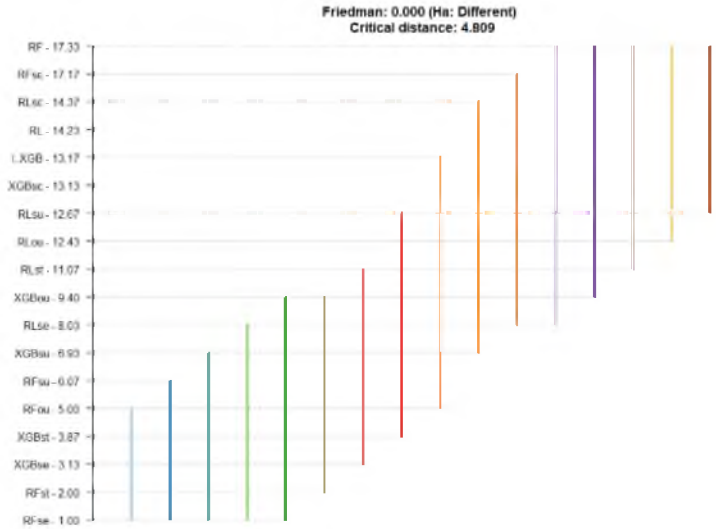
Figura 5.1 – *Comparativo dos modelos – box-plot da AUC*

Na literatura existem vários testes estatísticos utilizados para comparar algoritmos de aprendizagem estatísticos. Demsar [99] revisou o tema, recomendando testes não paramétricos simples, mas seguros e robustos para fazer comparações estatísticas de classificadores: a) teste *Wilcoxon*, para comparar dois classificadores, e; b) teste de Friedman, para comparação de mais classificadores entre si em vários conjuntos de dados, sendo que os resultados deste último também podem ser apresentados de forma organizada com os diagramas de diferença crítica, teste post-hoc de *Nemenyi*. Em seu trabalho[99], foi destacado que testes de análise de variância como ANOVA, que se baseia em suposições como: as amostras são aleatórias e independentes; as distribuições dos resíduos são normais; e, a variância dos dados nos grupos deve ser igual, provavelmente podem ser violadas ao analisar o desempenho de algoritmos de aprendizado de máquina. As violações dessas suposições têm um efeito ainda maior nos testes post-hoc, por exemplo *Tukey*, não sendo recomendado pelo autor[99].

No contexto deste estudo, foi utilizado o método *Friedman-Nemenyi* para testar a significância das diferenças entre médias múltiplas de ROC AUC dos modelos, valores médios de 30 amostras geradas no *cross-validation*.

A Figura 5.2 resume os resultados das estatísticas de Friedman, informando que há diferença entre os dados e modelos e o ranking dos algoritmos (quanto menor melhor) O modelo *Random Forest* balanceado pelo método SMOTE e ENN (RFse) obteve a primeira colocação por atingir a melhor média de ROC AUC na classificação dos

municípios. Na última colocação ficou modelo *Random Forest* base desbalanceada (RF), que obteve a pior média de ROC AUC. Entretanto, é necessário testar estatisticamente para obter uma maior exatidão nessa conclusão e verificar se há realmente uma diferença. O indicador CD (*Critical Distance*) informa se existe diferença estatística entre dois modelos, quando aplicados em um determinado conjunto de dados. O parâmetro para avaliar se os modelos são estatisticamente diferentes, se dá quando a diferença entre os rankings de cada modelo é maior do que o valor de CD, por exemplo os modelos RFse, RFst, XGBse, XGBst e RFou são estatisticamente iguais.



Fonte: Autoria própria

Figura 5.2 – *Comparativo dos modelos treinados utilizando teste de Friedman-Nemenyi*

O resultado apresentado no teste por si só não significa que os modelos melhores ranqueados são os modelos "campeões" e recomendados para implementação e uso. A avaliação mais importante acontece na próxima fase, em que os modelos são aplicados (escorados) em uma base de dados, que não foi utilizada no processo de treinamento.

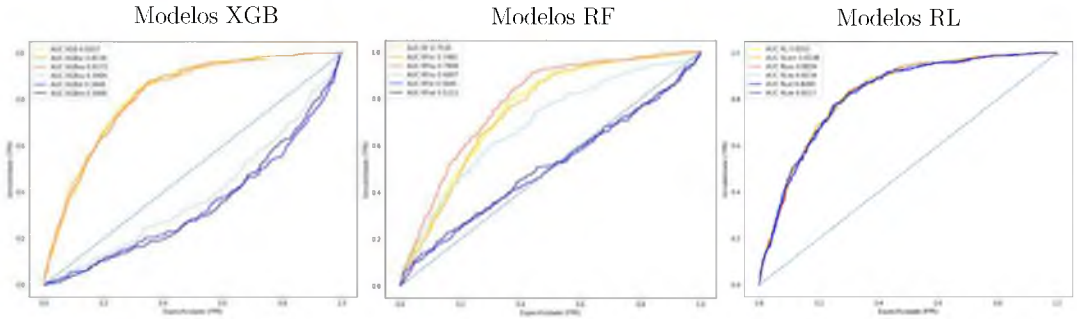
5.2. Teste fora da amostra de modelagem (*out-of-sample*)

Uma vez criados os modelos e aplicados na base de dado de treinamento, faz-se necessário avaliar se estes modelos geram resultados consistentes. Uma forma de avaliar a qualidade é aplicar os modelos desenvolvidos em uma base de dados que não foi utilizada na fase de modelagem. Nessa etapa, são analisadas a estabilidade dos modelos, verificando se estão "super ajustados" (*overfitting*), ou seja, quando o modelo apresenta excelentes resultados no treino, mas quando aplicado em uma base que não foi utilizada no processo de treinamento, é observada baixa performance. A seguir estão apresentados os resultados da aplicação dos modelos desenvolvidos na base de dados criada para esse teste (data-base de 2018) com descumprimentos observados ao longo de 2019.

Por meio da curva ROC AUC (Figura 5.3), é possível observar como foi o desempenho dos modelos. De maneira geral os modelos de RL tiveram desempenho semelhante ao observado no treinamento, com ligeiro aumento no valor de AUC (RL, de 0,7982 para 0,8253; RLsc, de 0,7981 para 0,8238; RLou, de 0,7970 para 0,8224; RLsu, de 0,7952 para 0,8234; RLst de 0,7953 para 0,8203; RLse, de 0,7963 para 08227).

Os modelos XGB e RF apresentaram para os modelos desenvolvidos com base nos métodos de amostragem balanceada: (i) SMOTE e *RandomUnderSampler*, (ii) SMOTE e *TomekLink* e (iii) SMOTE e ENN, resultados que confirmaram a situação de *overffiting*. Acredita-se que esse desempenho possa ter acontecido por conta do processo de *cross-validation* gerar repetição dos eventos da classe minoritária (*default*) nas 30 amostras, e também pelo grau de importância das variáveis para os modelos conter variáveis que não estariam com o mesmo nível de importância do observado na base de teste.

Analisando os demais modelos XGB, verifica-se que houve uma pequena queda de desempenho (XGB, de 0,8303 para 0,8207; XGBsc, de 0,8426 para 0,8178; XGBou, de 0,8308 para 0,8172), já os modelos RF tiveram valores bem abaixo do observado no treinamento e apresentou também certo nível de super ajuste (RF, de 0,9982 para 0,7530; RFsc, de 0,9953 para 07481; RFou de 0,9888 para 0,7848).



Fonte: Autoria própria

Figura 5.3 *– Curvas ROC AUC dos Modelos XGB, RF e RL aplicados na base de treino*

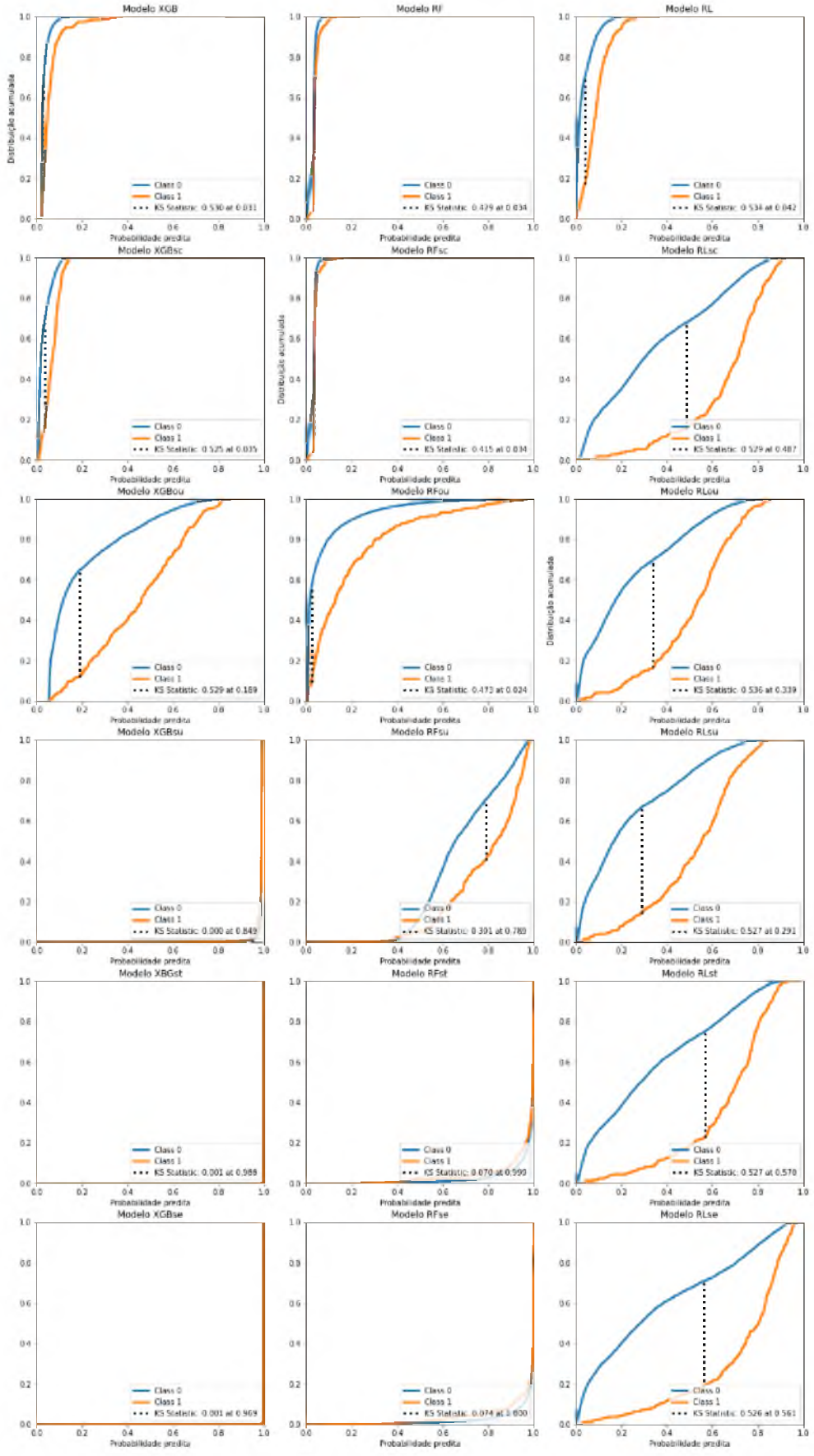
O comportamento dos modelos no indicador KS (Figura 5.4) também se apresentou de forma semelhante ao observado na fase de treinamento, com exceção, dos modelos já citados nos comentários sobre a Curva ROC.

O indicador KS de três dos seis modelos *XGBoost* apresentaram valores semelhantes, exceto o modelo XGBsc que teve uma variação para menor de 3,85% (XGB, de 0,526 para 0,530; XGBsc, de 0,546 para 0,525; XGBou, de 0,528 para 0,529), evidenciando certa estabilidade na classificação dos municípios. Numa análise combinada ROC e KS, esses modelos mostram-se promissores.

Os modelos *Random Forest* tiveram os piores desempenhos quando comparados com os modelos XGB e RL, por apresentarem comportamento de super ajuste. Dentre

eles o RFou ficou com o melhor desempenho de KS (treino: 0,921, teste: 0,473), mas abaixo dos demais modelos XGB e RL. Analisando os resultados até o momento de ROC e KS, os modelos gerados por meio do algoritmo RF não estariam entre os modelos escolhidos como "melhores".

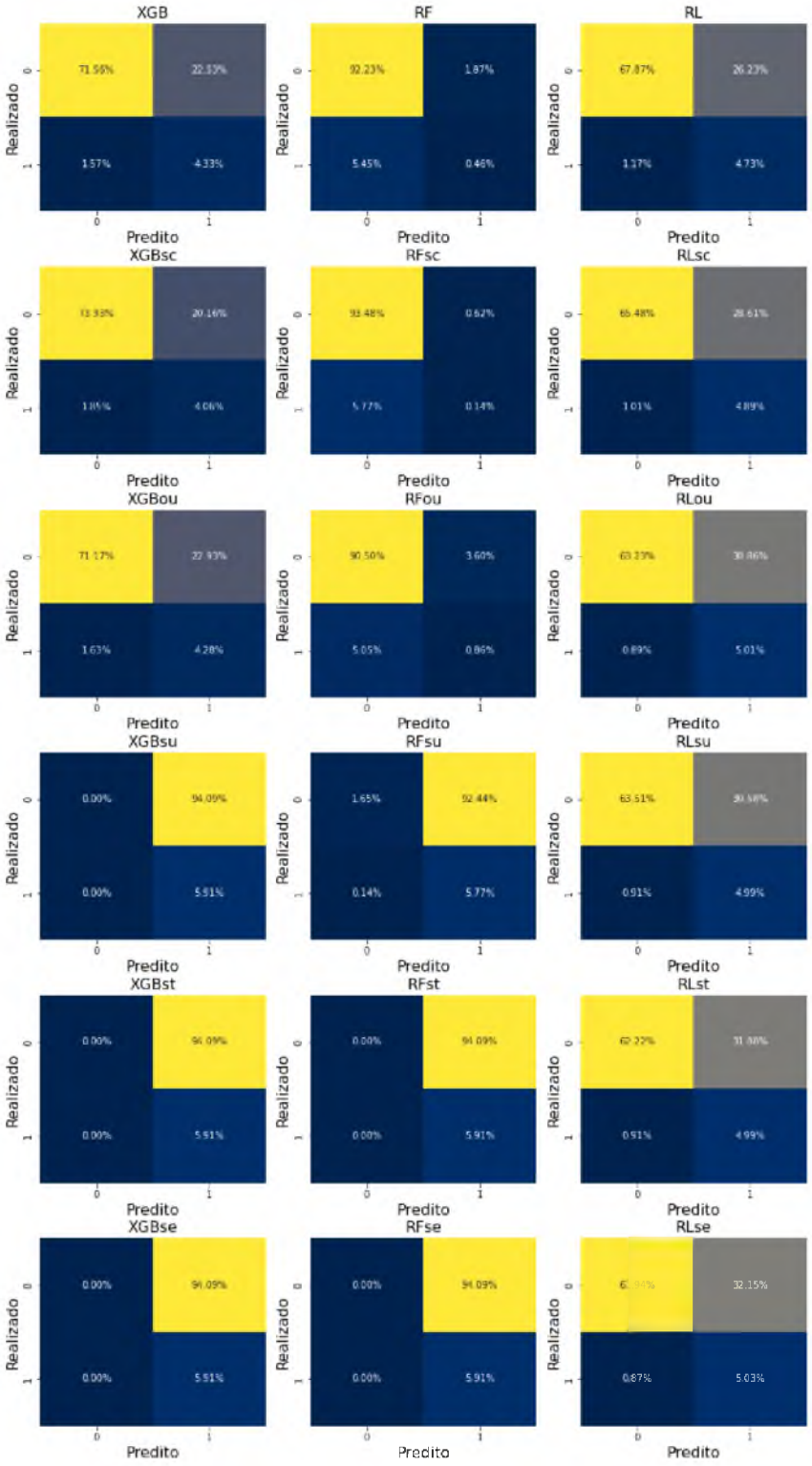
Já os modelos desenvolvidos com base na RL tiveram resultados de KS melhores dos que os apresentados no treinamento (RL, de 0,478 para 0,534; RLsc, de 0,485 para 0,529; RLou, de 0,481 para 0,536; RLsu, de 0,479 para 0,527; RLst de 0,481 para 0,527; RLse, de 0,489 para 0,526). A princípio esse comportamento também observado na Curva ROC AUC, pode ser explicado pela frequência maior de eventos de *default* observados na base de teste (5,91%), frente os *default*s na base de treinamento (3,77%). Há que se destacar a estabilidade dos coeficientes das variáveis dos modelos RL, mesmo sendo treinado em amostras diferentes.



Fonte: Autoria própria

Figura 5.4 – *Curvas KS dos Modelos XGB. RF e RL aplicados na base de treino*

As matrizes de confusão geradas com base na amostra de teste foram calculadas utilizando o ponto de corte otimizado (*G-mean*) observado no treinamento e apresentaram os seguintes comportamentos de percentuais de taxas de acerto e erros tipo I e II para cada modelo (Figura 5.5).



Fonte: Autoria própria

Figura 5.5 – *Matrizes de confusão dos Modelos XGB, RF e RL aplicados na base de teste*

Os resultados observados nas matrizes de confusão para os modelos *XGBoost* (XGB, XGBsc e XGBou) assim como encontrado na etapa de treinamento tiveram baixos percentuais de erro Tipo I (falso positivo), todavia os valores de erro Tipo II (falso negativo) ficaram proporcionalmente elevados, colocando "bons" municípios como "ruins". Nesse aspecto, os modelos valorizam uma classificação mais precisa dos municípios "ruins" em detrimento daqueles entes considerados "bons".

Os modelos *Random Forest* (RF, RFsc e RFou) tiveram comportamento contrário, minimizando o erro Tipo II, levando a uma estratégia não desejada para uma instituição financeira que é a de trazer para dentro da carteira potenciais municípios que poderiam gerar maior inadimplência, por conseguinte mais prejuízo.

Os modelos Regressão Logística, assim como os *XGBoost*, "pesaram a mão" para minimizar o erro Tipo I, gerando proporcionalmente mais erro Tipo II do que os modelos *XGBoost*. De fato, os modelos RL possuem característica mais conservadoras em termos de aceitação de municípios para a carteira.

No contexto de uma instituição financeira, a decisão de evitar prejuízos decorrentes de negócios com clientes ruins tende a pesar mais, sobretudo quando a rentabilidade trazida pelos negócios gerados com esses cliente seja considerada baixa ou baixíssima, nesse ponto uma avaliação criteriosa entre o custo em se perder negócios com bons clientes, "barrados" pelos modelos e o custo de trazer um cliente ruim, não detectado pelo modelo, deve ser mensurada pela instituição, ponderando outros aspectos de relacionamento com os municípios. Segundo o que é observado na prática das instituições, essa avaliação é muito difícil de se realizar [70].

A Tabela 5.1 e Tabela 5.2 a seguir resumem os resultados dos indicadores de desempenho de cada modelo aplicado na base de teste, tendo como referência os limites otimizados da Curva ROC AUC (*G-mean*), exceto o indicador KS, observados no treinamento.

Tabela 5.1: *Desempenho dos modelos no teste – Base desbalanceada e Sensível ao Custo*

	Base desbalanceada			Sensível ao Custo		
	XGB	RF	RL	XGBsc	RFsc	RLsc
AUC	0,747	0,529	0,761	0,736	0,509	0,762
Acurácia	0,759	0,927	0,726	0,780	0,936	0,704
F1	0,265	0,111	0,257	0,269	0,042	0,248
Precisão	0,161	0,197	0,153	0,167	0,184	0,146
Recall	0,734	0,077	0,801	0,687	0,024	0,828
KS	0,530	0,429	0,534	0,525	0,415	0,529

Legenda: XGB – *XGBoost*; RF – *Random Forest*; RL – Regressão Logística

Fonte: Autoria própria

Tabela 5.2: *Desempenho dos modelos no teste – Bases balanceada*

	RandomUndersample e *RandomOversample*			SMOTE e *RandomUndersample*			SMOTE e *Tomek Link*			SMOTE e *Edited Nearest Neighbors*		
	XGBou	RFou	RLou	XGBsu*	RFsu	RLsu	XGBst*	RFst*	RLst	XGBse*	RFse*	RLse
AUC	0,740	0,533	0,760	0,500	0,497	0,760	0,500	0,500	0,753	0,500	0,500	0,755
Acurácia	0,754	0,914	0,682	0,059	0,074	0,685	0,059	0,059	0,672	0,059	0,059	0,670
F1	0,258	0,165	0,240	0,112	0,111	0,241	0,112	0,112	0,233	0,112	0,112	0,234
Precisão	0,157	0,192	0,140	0,059	0,059	0,140	0,059	0,059	0,135	0,059	0,059	0,135
Recall	0,724	0,145	0,848	1,000	0,976	0,845	1,000	1,000	0,845	1,000	1,000	0,852
KS	0,529	0,473	0,536	0,000	0,301	0,527	0,001	0,070	0,527	0,001	0,074	0,526

Legenda: XGB – *XGBoost*; RF – *Random Forest*; RL – Regressão Logística

(*) alguns valores estão truncados considerando o desempenho na curva ROC AUC bem abaixo de 0,5.

Fonte: Autoria própria

Inicialmente, observamos que os resultados dos modelos RF, quando não apresentaram *overfitting*, tiveram valores de acurácia elevados frente os desenvolvidos com base no XGB e RL. Cabe destacar que embora esse indicador seja amplamente utilizado, como medida de precisão da classificação ele é quase universalmente inadequado, quando se trata de avaliar amostras com classes fortemente desbalanceadas. Esse comportamento se dá pelo fato de que uma alta acurácia (ou baixo erro) é alcançável por um modelo que prediz a classificação dos eventos somente na classe majoritária [59] [57]. Reforçando o ponto, Galar et al. [60] observam que em um conjunto de dados com classes desequilibradas, a acurácia não é mais uma medida adequada, uma vez que não distingue entre o número de eventos corretamente classificados em diferentes classes, levando a conclusões errôneas.

Os demais modelos XGB, excetuando os que ficaram super-ajustados, e os RLs apresentaram resultados aceitáveis, quanto à performance e comportamento tanto no contexto de treinamento quanto no teste, mantendo características já observadas nos tópicos anteriores.

5.3. Aplicação do modelo de classificação

Nesta etapa, é apresentada uma alternativa de classificação do risco de crédito dos municípios a partir de dois modelos escolhidos, utilizando uma escala de rating elaborada com base na distribuição das probabilidades estimadas dos algoritmos aplicados na base de teste.

Para a construção da escala de rating, segundo Sicsú [70] deve-se levar em consideração que para os primeiros modelos a serem implementados por uma instituição, um número grande de níveis de risco não é apropriado, a recomendação é de que uma escala de risco com 8 a 12 classes seja suficiente. Sendo que para amostras de 2 mil a 5 mil indivíduos, é possível trabalhar com seis a dez classes de modo a ter pelo menos cerca de 300 indivíduos em cada classe.

A escala deve apresentar, nas classes extremas, maiores níveis de observações. Outro ponto importante é observar se à medida que as probabilidades crescem, as frequências de "bons" aumentam e as de "ruins" diminuem. Esse é um sinal de que o modelo apresenta resultados coerentes [70]. A partir dos pontos comentados, adotou-se uma escala inicial de dez níveis de risco para os modelos escolhidos.

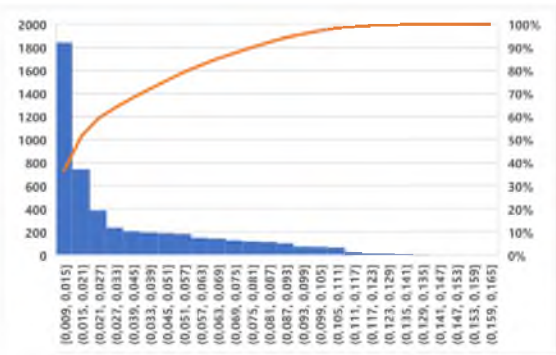
Diante do observados até o momento, sopesando os resultados da Curva ROC AUC, KS, indicadores *F1-score*, Precisão, *Recall*, Matriz de Confusão, foram escolhidos dois modelos, como opção para classificar os municípios na visão de risco de crédito, entre os desenvolvidos com *XGBoost* e Regressão Logística: XGBsc e RLsc.

Cabe destacar que os três modelos *XGBoost* (XBG, XGBsc e XGBou) tiveram desempenho semelhante nos indicadores, além de apresentarem pequenas diferenças entre o observado no treinamento e no teste. No caso da Regressão Logística, não houve diferenças significativas entre os modelos desenvolvidos, dentre as seis alternativas apresentadas.

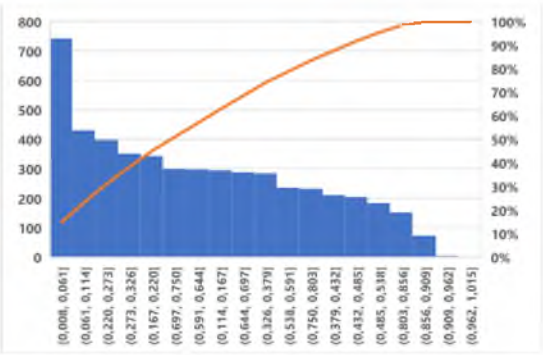
Os dois modelos escolhidos possuem características distintas, na forma de se chegar na probabilidade de *default*, tanto no algoritmo em essência quanto na importância das variáveis escolhidas para explicar a *target*. Essa diversidade é interessante para o gestor de riscos, que está responsável pelo gerenciamento da carteira de negócios com os municípios, pois a depender da estratégia a ser utilizada haverá uma alternativa de modelo mais adequada. Por exemplo, caso no momento o apetite a risco da instituição é correr menos riscos, reforçando um perfil mais conservador, o modelo RL pode ser uma boa opção.

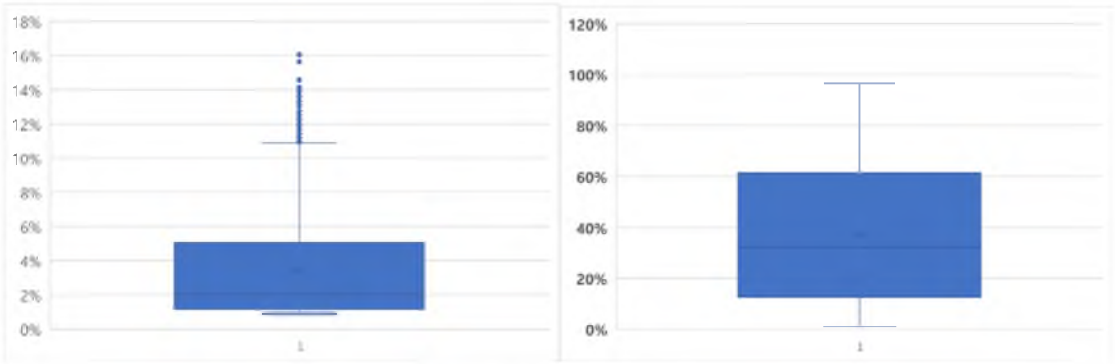
As distribuições de probabilidades acumuladas, juntamente com as matrizes de confusão, trazem essa sensibilidade quanto à "propensão" do modelo em classificar os municípios. Na Figura 5.6 estão apresentados graficamente as probabilidades de *default* estimadas para os entes públicos na base de teste, calculadas pelos modelos XBGsc e RLsc.

Modelo XGBsc



Modelo RLsc





Fonte: Autoria própria

Figura 5.6 – *Distribuição de probabilidades de default Modelos XGBsc e RLsc*

Particularmente, para o modelo RL as probabilidades estimadas tiveram valores elevados, tendo em vista a atribuição de pesos (custo) para as classes (0,1), dado o desbalanceamento. Nesse caso, quando for realizar a implementação do modelo em produção na instituição, recomenda-se que seja realizado ajuste para que a proporção entre "bons" e "ruins" da base de teste seja reconhecida. Para ajustar, há possibilidade, por exemplo, de calcular novo intercepto (β_0) da função logística, considerando a taxa de desbalanceamento das classes observadas na base de teste, utilizando a Equação (5.1 [70].

$$\beta_0^{corrigido} = \beta_0^{original} + ln\left(\frac{\pi_b}{\pi_r} \times \frac{n_r}{n_b}\right) \tag{5.1}$$

onde:

$\beta_0^{corrigido}$ = intercepto corrigido

$\beta_0^{original}$ = intercepto original da base de dados

π_b = probabilidade a priori de ser "bom"

π_r = probabilidade a priori de ser "ruim"

n_b = tamanho da amostra original de "bons"

n_r = tamanho da amostra original de "ruins"

A Tabela 5.3 a seguir dispõe, para cada modelo, as frequências de probabilidades acumuladas de *default* dos municípios classificados em cada nível de risco (10), considerando os eventos das classes "bons" e "ruins" observadas na base de teste, além da proporção de "bons" e "ruins" (B/R) em rating da escala de risco. Faz-se importante esclarecer que a distribuição das observações nos ratings foi feita com base na própria característica das distribuições de probabilidade observada, portanto as escalas são distintas dado o modelo de risco. Para efeito deste estudo, não foi realizado o ajuste do intercepto da função logística para o modelo RLsc.

Tabela 5.3: *Municípios "bons" e "ruins" por nível risco de risco de crédito*

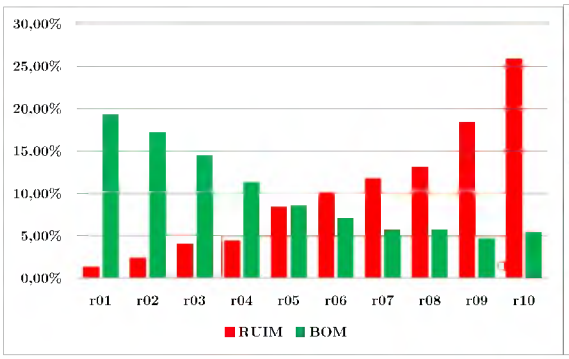
Risco	Modelo XGBsc				Modelo RLsc			
	Intervalo (PD)	Ruim	Bom	B/R	Intervalo (PD)	Ruim	Bom	B/R
r01	- 1,00%	1,35%	19,4%	14,39	- 5,00%	0,67%	12,1%	17,95
r02	1,00% 1,40%	2,36%	17,3%	7,33	5,00% 20,00%	2,69%	22,9%	8,50
r03	1,40% 1,90%	4,04%	14,4%	3,57	20,00% 30,00%	2,02%	15,2%	7,54
r04	1,90% 2,70%	4,38%	11,3%	2,59	30,00% 40,00%	6,06%	11,2%	1,85
r05	2,70% 3,90%	8,42%	8,6%	1,02	40,00% 50,00%	5,39%	7,3%	1,35
r06	3,90% 5,00%	10,10%	7,1%	0,70	50,00% 58,00%	7,41%	6,4%	0,86
r07	5,00% 6,10%	11,78%	5,7%	0,49	58,00% 64,00%	9,09%	6,4%	0,70
r08	6,10% 7,60%	13,13%	6,2%	0,47	64,00% 70,00%	14,81%	6,0%	0,41
r09	7,60% 9,00%	18,52%	4,6%	0,25	70,00% 80,00%	31,65%	8,7%	0,27
r10	9,00% 100,00%	25,93%	5,4%	0,21	80,00% 100,00%	20,20%	3,9%	0,19

Legenda: PD – Probabilidade de *Default*

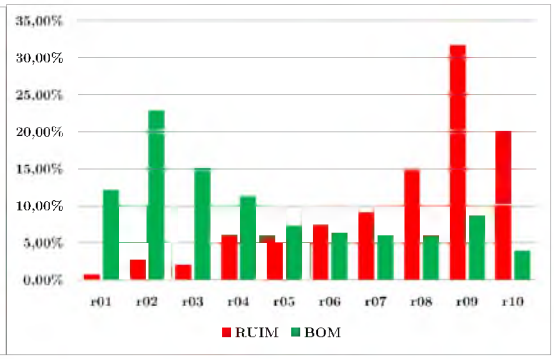
Fonte: Autoria própria

Graficamente a classificação dos municípios nos níveis de risco ficou assim distribuída, conforme a Figura 5.7 a seguir:

Modelo XGBsc



Modelo RLsc



Fonte: Autoria própria

Figura 5.7 *– Distribuição de Municípios "bons" e "ruins"*

No processo de implementação, a instituição pode optar por uma escala mais ou menos granular, a depender de como está estruturada, observando sobretudo o processo decisório associado, já que os modelos em si devem "se encaixar" na dinâmica instituída pela IF. Outro ponto importante que pode ser adotado a partir da escala de risco é o estabelecimento de um ponto de corte sobre aceitar ou não um determinado crédito com o município, a escolha de certa maneira leva em consideração a relação risco X retorno, no contexto da gestão de portfólio.

5.3.1. Comparação – Modelos x CAPAG

Uma dúvida que pode ser gerada é se esses modelos desenvolvidos estão coerentes sob uma ótica de classificação de risco já instituída e publicamente divulgada, como por exemplo a CAPAG. Relembrando, a CAPAG é uma avaliação da capacidade de pagamento dos entes públicos calculada a partir de três indicadores gerados com base nas

informações econômico-financeiras. Esses indicadores combinados são traduzidos em quatro níveis de risco ("A", "B", "C" e "D") numa escala de melhor ("A") para pior ("D").

Para realizarmos a comparação entre os resultados dos modelos e a avaliação da CAPAG, faz necessário que as informações estejam em uma mesma data-base de referência. No caso desta pesquisa, foram utilizadas as informações econômico-financeiras dos municípios compiladas nos indicadores IFGF da FIRJAN calculados com base no ano de 2018 (base de teste). Essas informações, segundo relatório metodológico da FIRJAN [83], são inteiramente construídas com base em resultados fiscais oficiais, declarados pelas próprias prefeituras. Conforme estabelecido pelo Artigo 51, da Lei de Responsabilidade Fiscal [9], os municípios devem encaminhar suas contas para a Secretaria do Tesouro Nacional (STN), até o dia 30 de abril do ano seguinte ao exercício de referência, a partir quando o órgão dispõe de 60 dias para disponibilizá-las ao público, por meio do Sistema de Informações Contábeis e Fiscais do Setor Público Brasileiro (Siconfi). No caso específico da base de 2018, a data final de consolidação do banco de dados do IFGF foi o dia 14 de julho de 2019, conforme informado no relatório da metodologia[83].

As informações da CAPAG também seguem a mesma dinâmica de cronograma e fonte de dados, já que essas informações possuem origem única (Siconfi). A base de dados utilizada da CAPAG, foi uma extração realizada em agosto de 2019, contendo as informações fechadas dos municípios referentes ao ano base de 2018. Assim torna-se possível realizar a comparação entre as duas formas de avaliação de riscos dos municípios.

Outro ponto importante a destacar é que dentre os 5.039 municípios classificados pelos modelos XGBsc e RLsc, 1.775 não possuem a classificação de risco da CAPAG na base de dados de 2018, ou seja esses entes estão identificados com o rótulo "n.d." (não disponível). A Tabela 5.4 e Tabela 5.5 a seguir trazem uma matriz de riscos entre os modelos desenvolvidos e a metodologia CAPAG.

Tabela 5.4: *Comparativo modelo XGBsc X CAPAG*

XGBsc	**CAPAG**					**TOTAL**
Risco	**A**	**B**	**C**	**D**	**n.d.**	
r01	273	173	202		273	921
r02	134	130	293		267	824
r03	73	112	312		198	695
r04	42	66	295		146	549
r05	35	37	208	1	150	431
r06	18	46	150		151	365
r07	11	28	125		142	306
r08	10	25	142	2	152	331
r09	5	18	128		122	273
r10	3	12	145		174	334
TOTAL	**604**	**647**	**2000**	**3**	**1775**	**5029**

Fonte: Autoria própria

Tabela 5.5: *Comparativo modelo RLsc X CAPAG*

RLsc	CAPAG					TOTAL
Risco	A	B	C	D	n.d.	
r01	149	119	127		179	574
r02	232	177	354		328	1091
r03	76	118	323		210	727
r04	44	65	297		143	549
r05	31	31	168	1	129	360
r06	21	35	132		135	323
r07	22	33	132		141	328
r08	12	28	145	1	142	328
r09	12	27	231		234	504
r10	5	14	91	1	134	245
TOTAL	**604**	**647**	**2000**	**3**	**1775**	**5029**

Fonte: Autoria própria

Antes de avançar é importante esclarecer que se tratam de formas totalmente diferentes de avaliação do risco de crédito dos municípios, ou seja, a comparação "fria" entre elas pode gerar conclusões equivocadas. Isto posto, comparando os modelos com a CAPAG, verifica-se que eles possuem, de maneira geral, uma certa lógica quanto à concentração dos municípios nos riscos. Não se observa uma inversão forte em relação à classificação, por exemplo os municípios mais bem classificados pela CAPAG encontram-se distribuídos nos melhores níveis de risco dos modelos, sendo possível concluir que ambos possuem um mesmo sentido de risco para a maioria dos municípios, obviamente calibrado em função das peculiaridades de cada método de mensuração.

Observa-se também, como comentado anteriormente, que o modelo RLsc possui um nível de conservadorismo maior, quanto à classificação de risco, do que o modelo XGBsc.

6. Considerações Finais

Este estudo apresentou alternativas para o desenvolvimento de modelo de mensuração do risco de crédito para municípios brasileiros. A partir da mineração de dados públicos e aplicação de métodos estatísticos e computacionais, foi possível estimar a probabilidade de *default* (inadimplência) do ente público.

Observou-se que o método computacional *XGBoost* (XGB) se mostrou superior ao *Random Forest*, considerando as bases de dados e o objetivo proposto. Quanto ao método estatístico tradicional, Regressão Logística (RL), constatou-se ser uma ferramenta robusta para estimar o risco de crédito, não sendo de maneira geral inferior ao método computacional. Ao final, os dois métodos (XGB e RL) mostraram-se alternativas viáveis a serem aplicadas no processo de gerenciamento do risco de crédito em portfólios de negócios com municípios, dadas as suas características e peculiaridades, podendo ser acionados a depender da estratégia de gestão dos riscos e negócios da instituição em determinado momento.

Os dados abertos utilizados na modelagem se mostraram suficientes para desenvolvimento dos modelos. A definição da variável *target* a ser prevista, no caso *default*, é uma informação particular de cada instituição financeira (IF), sendo construída a partir do referencial do que seria considerado *default* para ela. Nesse sentido, a base utilizada neste estudo tem o viés, em certa medida, do que seria "bom" ou "ruim" para a instituição que forneceu os dados. Todavia, o processo pode ser replicado por outras instituições, sem prejuízo do racional empregado.

Essa forma de medir o risco de crédito dos municípios é uma medida alternativa voltada também para orientar o apetite ao risco que a instituição está disposta a assumir, nesse caso está vinculado a cada ente público, mas pode ser no nível de portfólio. No contexto de uma IF, essa dimensão é a mais importante sob o ponto de vista dos gestores, por ela representar a possibilidade de ocorrência de perdas nos negócios, com reflexos na redução da lucratividade e na possibilidade de impacto no capital próprio, podendo expor a IF ao desenquadramento dos requisitos mínimos de capital próprio [53].

Estão à disposição da instituição, os modelos desenvolvidos que tiveram melhores indicadores de qualidade (estatísticas, matriz de confusão, impactos no uso) para avaliação da viabilidade de implementação na organização.

Outros achados neste estudo podem ensejar novas pesquisas ou aprimoramentos futuros:

a) Resultados observados no comportamento das variáveis explicativas dos municípios, como por exemplo a concentração dos "melhores" municípios estimados pelos modelos nas regiões Sul e Sudeste.

b) Comportamento dos algoritmos utilizados quando aplicados nas amostras geradas pelas técnicas de amostragem balanceada, haja vista os resultados ficarem aquém das expectativas observadas em trabalhos citados no referencial teórico.

c) Os modelos voltados para mensurar o risco de crédito foram pouco sensíveis às variáveis sociais, levando a crer que aspectos como a desigualdade social sejam tratados de uma outra forma, caso o objetivo seja avaliar o ente público sob o aspecto de oportunidade de negócios, para fomentar o desenvolvimento e reduzir a desigualdade socioeconômica.

Do ponto de vista acadêmico, este trabalho contribui para a redução do gap científico relacionado ao uso de métodos computacionais e estatísticos para a estimativa da inadimplência municipal. Seu uso não se restringe a instituições financeiras, podendo ser insumo, por exemplo, para adoção de políticas públicas voltadas para a melhoria dos índices de desenvolvimento econômico, sendo uma variável importante para a decisão de financiamento de projetos municipais capazes de contribuir para a melhoria da qualidade de vida da população em geral.

7. Referências Bibliográficas

[1] K. Benetti, "Challenges of corporate risk management after global financial crisis," in *The 10th International Days of Statistics and Economics*, 2016, no. 2014, pp. 154–162.

[2] T. K. Quon, D. Zeghal, and M. Maingot, "Enterprise Risk Management and Firm Performance," *Procedia - Soc. Behav. Sci.*, vol. 62, pp. 263–267, 2012.

[3] S. V. Jerić and M. Primorac, "Data mining for assessing the credit risk of local government units in Croatia," *Croat. Oper. Res. Rev.*, vol. 8, no. 1, pp. 193–205, 2017.

[4] Basel Committee on Banking Supervision - BCBS, *Basel III: A global regulatory framework for more resilient banks and banking systems*, Rev.June. Basel: Bank for International Settlements, 2011.

[5] Basel Committee on Banking Supervision - BCBS, *Basel II: International Convergence of Capital Measurement and Capital Standards - A Revised Framework & Comprehensive Version*, no. June. Basel, Switzerland, 2006.

[6] CMN, "Resolução nº 4.557," *Banco Central do Brasil*. BACEN, Brasília, pp. 1–32, 2017.

[7] M. F. Hellwig, "Systemic Risk in Financial Sector: An Analysis of the Subprimi-Mortgage financial Crisis," *Economist (Leiden).*, no. 2, pp. 129–207, 2009.

[8] República Federativa do Brasil, *Constituição da República Federativa do Brasil: texto constitucional promulgado em 05.10.1988, com as alterações adotadas pelas Emendas constitucionais nº 1/92 a 110/2021, pelo Decreto legislativo nº 186/2008 e pelas Emendas const. de revisão nº 1 a 6/94*, V. 57ª 2021. Brasília (DF): Edições Câmara, 1988.

[9] República Federativa do Brasil, *Lei de Responsabilidade Fiscal - LRF: Lei Complementar nº 101, de 4 de maio de 2000*, 4ª reimpre. Brasília (DF), 2000.

[10] República Federativa do Brasil, "Resolução nº 40." Senado Federal, Brasília (DF), p. 3, 2001.

[11] República Federativa do Brasil, "Resolução nº 43." Senado Federal, Brasília (DF), p. 20, 2001.

[12] República Federativa do Brasil, "Resolução nº 48." Senado Federal, Brasília (DF), p. 6, 2007.

[13] FNP Frente Nacional de Prefeitos, "Multi Cidades - Dinancas dos Municípios do Brasil," Vitória (ES), 2021.

[14] Banco Mundial, "COVID-19 no Brasil: Impactos e Respostas de Políticas Públicas.," 2020.

[15] Ibope, "Impactos da COVID-19 nos municípios." Ibope Inteligência, São Paulo, p. 52, 2020.

[16] A. Hamerle, A. Dartsch, R. Jobst, and K. Plank, "Integrating macroeconomic risk factors into credit portfolio models," *J. RISK Model Valid.*, vol. 5, no. 2, pp. 3–24, 2011.

[17] S. Cohen, M. Doumpos, E. Neofytou, and C. Zopounidis, "Assessing financial distress where bankruptcy is not an option: An alternative approach for local municipalities," *Eur. J. Oper. Res.*, vol. 218, no. 1, pp. 270–279, 2012.

[18] D. Alaminos, S. M. Fernández, F. García, and M. A. Fernández, "Data mining for municipal financial distress prediction," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10933 LNAI, no. July, pp. 296–308, 2018.

[19] D. Buendía-Carrillo, J. Lara-Rubio, A. Navarro-Galera, and M. E. Gómez-Miranda, "The impact of population size on the risk of local government default," *Int. Tax Public Financ.*, vol. 27, no. 5, pp. 1264–1286, 2020.

[20] E. Gorina, C. Maher, and M. Joffe, "Local Fiscal Distress: Measurement and Prediction," *Public Budg. Financ.*, vol. 38, no. 1, pp. 72–94, 2018.

[21] E. Galariotis, A. Guyot, M. Doumpos, and C. Zopounidis, "A novel multi-attribute benchmarking approach for assessing the financial performance of local governments: Empirical evidence from France," *Eur. J. Oper. Res.*, vol. 248, no. 1, pp. 301–317, 2016.

[22] J. Saary-Littman, M. A. Banks, and S. Leung, "Principles and Practices Benchmarking Survey - Credit Portfolio Management at the Tail End of the Credit Cycle," IACPM - International Association of Credit Portfolio Managers, New York, 2019.

[23] A. Melo Mariano and M. Rocha Santos, "Revisão da Literatura: Apresentação de uma Abordagem Integradora Structural Equations View project Service Quality View project," in *XXVI Congreso Internacional de la Academia Europea de Dirección y Economía de la Empresa (AEDEM)*, 2017, no. September, p. v.26.

[24] J. CAPECI, "CREDIT RISK, CREDIT RATINGS, AND MUNICIPAL BOND YIELDS - A PANEL STUDY," *Natl. Tax J.*, vol. 44, no. 4, 1, pp. 41–56, 1991.

[25] C. Li, Y. Xie, and J. Zhou, "National Level, City Level Auditor Industry Specialization and Cost of Debt," *Account. HORIZONS*, vol. 24, no. 3, pp. 395–417, 2010.

[26] P. Hajek, "Municipal credit rating modelling by neural networks," *Decis. Support Syst.*, vol. 51, no. 1, pp. 108–118, 2011.

[27] S. P. Davies, L. E. Johnson, and S. Lowensohn, "Ambient Influences on Municipal Net Assets: Evidence from Panel Data," *Contemp. Account. Res.*, vol. 34, no. 2,

pp. 1156–1177, 2017.

[28] A. W. Beck, "Opportunistic financial reporting around municipal bond issues," *Rev. Account. Stud.*, vol. 23, no. 3, pp. 785–826, 2018.

[29] E. Padovani, D. W. Young, and E. Scorsone, "The role of a municipality's financial health in a firm's siting decision," *Bus. Horiz.*, vol. 61, no. 2, pp. 181–190, 2018.

[30] E. Padovani, L. Rescigno, and J. Ceccatelli, "Municipal Bond Debt and Sustainability in a Non-Mature Financial Market: The Case of Italy," *SUSTAINABILITY*, vol. 10, no. 9, 2018.

[31] M. Adelino, I. Cunha, and M. A. Ferreira, "The Economic Effects of Public Financing: Evidence from Municipal Bond Ratings Recalibration," *Rev. Financ. Stud.*, vol. 30, no. 9, pp. 3223–3268, 2017.

[32] J. Cornaggia, K. J. Cornaggia, and R. D. Israelsen, "Credit Ratings and the Cost of Municipal Financing," *Rev. Financ. Stud.*, vol. 31, no. 6, pp. 2038–2079, Jun. 2018.

[33] M. Painter, "An inconvenient cost: The effects of climate change on municipal bonds," *J. financ. econ.*, vol. 135, no. 2, pp. 468–482, 2020.

[34] J. Wang, C. Wu, and F. X. Zhang, "Liquidity, default, taxes, and yields on municipal bonds," *J. Bank. \& Financ.*, vol. 32, no. 6, pp. 1133–1149, Jun. 2008.

[35] M. Schwert, "Municipal Bond Liquidity and Default Risk," *J. Finance*, vol. 72, no. 4, pp. 1683–1721, 2017.

[36] G. Purdy, "ISO 31000:2009 - Setting a new standard for risk management: Perspective," *Risk Anal.*, vol. 30, no. 6, pp. 881–886, 2010.

[37] ABNT, "NBR ISO/IEC 31010. Técnicas para o Processo de Avaliação de Riscos," Rio de Janeiro (RJ), 2012.

[38] ABNT, "ABNT NBR ISO 31000:2009." Rio de Janeiro, p. 27, 2009.

[39] C. Lalonde and O. Boiral, "Managing risks through ISO 31000: A critical analysis," *Risk Manag.*, vol. 14, no. 4, pp. 272–300, 2012.

[40] G. Purdy, "ISO 31000:2009 - Setting a new standard for risk management: Perspective," *Risk Anal.*, vol. 30, no. 6, pp. 881–886, 2010.

[41] P. Chapman *et al.*, *CRISP-DM 1.0 Step-by-step data mining guide*. IBM, 2000.

[42] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 526–534, 2021.

[43] Basel Committee on Banking Supervision - BCBS, *International Convergence of Capital Measurement and Capital Standards*. Basel, Switzerland, 1988.

[44] Basel Committee on Banking Supervision - BCBS, *The Amendment to the Capital Accord to Incorporate Market Risk*. Basel, Switzerland, 1996.

[45] M. Tiesset and P. Troussard, "Regulatory Capital and Economic Capital," *Financ. Stab. Rev. Banq. Fr.*, no. 7, pp. 59–74, 2005.

[46] Bacen, "Circular nº 3.648: Abordagem IRB - Sistemas Internos de Classificação do Risco de Crédito." Banco Central do Brasil, Brasília (DF), p. 120, 2013.

[47] E. I. Altman, "Applications of Distress Prediction Models: What Have We Learned After 50 Years from the Z-Score Models?," *Int. J. Financ. Stud.*, vol. 6, no. 3, p. 70, 2018.

[48] F. Barboza, H. Kimura, V. A. Sobreiro, and L. F. C. Basso, "Credit risk: From a systematic literature review to future directions," *Corp. Ownersh. Control*, vol. 13, no. 3continued2, pp. 326–346, 2016.

[49] IASB, "IFRS 9 Financial Instruments," London UK, 2014.

[50] Basel Committee on Banking Supervision - BCBS, *International Convergence of Capital Measurement and Capital Standards*. Basel: Bank for International Settlements, 2006.

[51] Z. Novotny-farkas, "The Interaction of the IFRS 9 Expected Loss Approach with Supervisory Rules and Implications for Financial Stability," *Account. Eur.*, vol. 13, pp. 197–227, 2016.

[52] D. Bholat, R. M. Lastra, S. M. Markose, A. Miglionico, and K. Sen, "Non--performing loans at the dawn of IFRS 9: regulatory and accounting treatment of asset quality," *J. Bank. Regul.*, vol. 19, no. (1), pp. 33–54, 2018.

[53] A. Michael, S. Venkat, Z. Mogul, S. Leung, M. A. Banks, and J. Saary-Littman, "Risk Appetite Frameworks: Insights into Evolving Global Practices," *Global Credit Review*, vol. 5. World Scientific Publishing Company and Risk Management Institute, NUS, pp. 1–17, 2015.

[54] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009.

[55] Y.-C. Chang, K.-H. Chang, H.-H. Chu, and L.-I. Tong, "Establishing decision tree-based short-term default credit risk assessment models," *Commun. Stat. METHODS*, vol. 45, no. 23, pp. 6803–6815, 2016.

[56] A. N. Tarekegn, M. Giacobini, and K. Michalak, "A review of methods for imbalanced multi-label classification," *Pattern Recognit.*, vol. 118, p. 107965, 2021.

[57] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1–50, 2016.

[58] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016.

[59] J. Brownlee, *Imbalanced Classification with Python - Choose Better Metrics, Balance Skewed Classes, and Apply Cost-Sensitive Learning*, V1.3. 2021.

[60] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *IEEE Trans. Syst. Man. Cybern.*, vol. 42, no. 4, pp. 463–484, 2012.

[61] G. E. A. P. A. Batista, A. L. C. Bazzan, and M. C. Monard, "Balancing Training Data for Automated Annotation of Keywords: a Case Study," *Proc. Second Brazilian Work. Bioinforma.*, no. January, pp. 35–43, 2003.

[62] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[63] N. Thai-Nghe, Z. Gantner, and L. Schmidt-Thieme, "Cost-sensitive learning methods for imbalanced data," in *Proceedings of the International Joint Conference on Neural Networks*, 2010.

[64] P. Turney, "Types of Cost in Inductive Concept Learning," in *Procedures of the Cost-Sensitive Learning Workshop at the 17th ICML-2000 Conference*, 2000, p. 7.

[65] M. Leo, S. Sharma, and K. Maddulety, "Machine learning in banking risk management: A literature review," *Risks*, vol. 7, no. 1, 2019.

[66] D. Andriosopoulos, M. Doumpos, P. M. Pardalos, and C. Zopounidis, "Computational approaches and data analytics in financial services: A literature review," *J. Oper. Res. Soc.*, vol. 70, no. 10, pp. 1581–1599, 2019.

[67] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," *Expert Syst. Appl.*, vol. 83, pp. 405–417, 2017.

[68] F. Butaru, Q. Chen, B. Clark, S. Das, A. W. Lo, and A. Siddique, "Risk and risk management in the credit card industry," *J. Bank. Financ.*, vol. 72, pp. 218–239, Nov. 2016.

[69] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3446–3453, 2012.

[70] A. L. Sicsú, *Credit Scoring - Desenvolvimento, Implantação, Acompanhamento*, 1ª. São Paulo: Blucher, 2010.

[71] L. Breiman, "Random Forest," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[72] C. Chen, A. Liaw, and L. Breiman, "Using Random Forest to Learn Imbalanced Data," *Discovery*, no. 1999, pp. 1–12, 2004.

[73] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining, 2016, vol. 13-17-Augu, pp. 785–794.

[74] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, 2005.

[75] S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 124–136, 2015.

[76] R. S. Wazlawick, *Metodologia de Pesquisa em Ciência da Computação*, 6ª Impress. Rio de Janeiro: Elsevier Editora Ltda., 2009.

[77] R. S. Wazlawick, "Capítulo 5 - Estilos de pesquisa correntes em computação," in *Metodologia De Pesquisa Para Ciência Da Computação*, Segunda Ed., R. S. Wazlawick, Ed. Rio de Janeiro: Elsevier Editora Ltda., 2009, pp. 5–15.

[78] V. Belton and T. J. Stewart, *Multiple Criteria Decision Analysis: An Integrated Approach*, 1 st. Massachusetts: Springer Science+Business Media, 2002.

[79] J. ribeirao S. Guimarães and P. D. M. Jannuzzi, "IDH, indicadores sintéticos e suas aplicações em políticas públicas," *R. B. Estud. URBANOS E Reg.*, vol. 7, N1, no. 2002, pp. 73–90, 2005.

[80] IPEA, *Atlas da Vulnerabilidade Social nos Municípios Brasileiros*. Brasília (DF): IPEA, 2015.

[81] Ministério da Fazenda, "Portaria nº 501, de 23 de novembro de 2017," *Imprensa Nacional*. Secretaria do Tesouro Nacional, Brasília (DF), pp. 1–6, 2017.

[82] Ministério da Fazenda, "Portaria nº 882, de 18 de dezembro de 2018," *Imprensa Nacional*. Secretaria do Tesouro Nacional, Brasília (DF), pp. 1–4, 2018.

[83] FIRJAN, "Metodologia - IFGF Índice Firjan Gestão Fiscal." Federação das Indústrias do Estado do Rio de Janeiro, Rio de Janeiro, pp. 1–6, 2019.

[84] Ministério da Fazenda, "Portaria MF nº 1, de 3 de janeiro de 2017," *Imprensa Nacional*. Ministério da Previdência Social, Brasília (DF), p. 2, 2017.

[85] Ministério da Previdência Social, "Portaria nº 402, De 10 De Dezembro De 2008," *Imprensa Nacional*. Brasília (DF), p. 29, 2008.

[86] Ministério da Fazenda, "Portaria nº10, de 08 de setembro de 2017.," *Imprensa Nacional*. Secretaria da Previdência, Brasília (DF), p. 19286818, 2017.

[87] Secretaria de Previdência, "Indicador de Situação Previdencária - ISP RPPS." Brasília (DF), pp. 1–51, 2018.

[88] República Federativa do Brasil, "Lei nº 9.717." Presidência da República do Brasil, Brasília (DF), p. 7, 1998.

[89] República Federativa do Brasil, "Lei nº 8.212." Presidência da República do Brasil, Brasília (DF), p. 37, 1991.

[90] República Federativa do Brasil, "Lei nº 8.213." Presidência da República do Brasil, Brasília (DF), p. 40, 1991.

[91] IBGE, "Pesquisa Nacional por Amostra de Domicílios Contínua - Notas Metodológicas." Instituto Brasileiro de Geografia e Estatística - IBGE, Rio de Janeiro (RJ), p. 47, 2014.

[92] D. J. Hand and W. E. Henley, "Statistical Classification Methods in Consumer Credit Scoring: a Review," *R. Stat. Soc.*, vol. 160, pp. 523–541, 1997.

[93] J. D. Sterman, "All models are wrong: Reflections on becoming a systems scientist," *Syst. Dyn. Rev.*, vol. 18, no. 4, pp. 501–531, 2002.

[94] J. Demšar *et al.*, "Orange: Data mining toolbox in python," *J. Mach. Learn. Res.*, vol. 14, no. October 2019, pp. 2349–2353, 2013.

[95] H. Liu, F. Hussain, C. L. Tan, and M. Dash, "Discretization: An Enabling Technique," *Data Min. Knowl. Discov.*, vol. 6, no. 4, pp. 393–423, 2002.

[96] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Comput. Stat. Data Anal.*, vol. 52, no. 4, pp. 2249–2260, 2008.

[97] H. He and Y. Ma, *Imbalanced Learning Fundations, Algorithms, and Applications*, 1ª. IEEE Press, 2013.

[98] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 694–699, 2002.

[99] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.

[100] CMN, "Resolução nº 4.553," *Banco Central do Brasil*. BACEN, Brasília (DF), p. 5, 2017.

8. Apêndice

Quadro 8.1: Atributos do Atlas da Vulnerabilidade Social nos Municípios e Regiões Metropolitanas Brasileiras

Fonte: http://ivs.ipea.gov.br/index.php/pt/biblioteca
Arquivo: IVS Municipal – Base completa

Nº	Sigla	Nome	Definição
1	ano	Ano	Ano de referência das informações
2	brasil	Brasil	Brasil
3	regiao	Macrorregião	Macrorregião
4	nome_regiao	Nome da macrorregião	Nome da macrorregião
5	uf	Código da Unidade da Federação	Código da Unidade da Federação
6	nome_uf	Nome da Unidade da Federação	Nome da Unidade da Federação
7	rm	Código da Região Metropolitana	Código da Região Metropolitana
8	nome_rm	Nome da Região Metropolitana	Nome da Região Metropolitana
9	municipio	Código do Município (7 dígitos)	Código utilizado pelo IBGE para identificação do município (com digito verificador).
10	Municipio_6digt	Código do Município (6 dígitos)	Código utilizado pelo IBGE para identificação do município.
11	nome_municipio	Nome do Município	Nome do Município
12	udh	Código da Unidade de Habitação	Código da Unidade de Habitação
13	nome_udh	Nome da Unidade de Habitação	Nome da Unidade de Habitação
14	ivs	IVS	Índice de Vulnerabilidade Social. Média aritmética dos índices das dimensões: IVS Infraestrutura Urbana, IVS Capital Humano e IVS Renda e Trabalho.
15	ivs_infraestrutura_urbana	IVS Infraestrutura Urbana	Índice da dimensão Infraestrutura Urbana, é um dos 3 índices que compõem o IVS. É obtido através da média ponderada de índices normalizados construídos a partir dos indicadores que compõem esta dimensão, a saber: 1) Percentual da população que vive em domicílios urbanos sem o serviço de coleta de lixo (peso: 0,300); 2) Percentual de pessoas em domicílios com abastecimento de água e esgotamento sanitário inadequados (peso: 0,300); 3) Percentual de pessoas em domicílios vulneráveis à pobreza e que gastam mais de uma hora até o trabalho no total de pessoas ocupadas,

			vulneráveis e que retornam diariamente do trabalho (peso: 0,400).
16	ivs capital hum ano	IVS Capital Humano	Índice da dimensão Capital Humano, é um dos 3 índices que compõem o IVS. Obtido através da média ponderada de índices normalizados construídos a partir dos indicadores que compõem esta dimensão, a saber: 1) Mortalidade até um ano de idade (peso: 0,125); 2) Percentual de crianças de 0 a 5 anos que não frequenta a escola (peso: 0,125); 3) Percentual de crianças de 6 a 14 anos que não frequenta a escola (peso: 0,125) ; 4) Percentual de mulheres de 10 a 17 anos de idade que tiveram filhos (peso: 0,125); 5) Percentual de mães chefes de família, sem fundamental completo e com pelo menos um filho menor de 15 anos de idade, no total de mães chefes de família (peso: 0,125); 6) Taxa de analfabetismo da população de 15 anos ou mais de idade (peso: 0,125); 7) Percentual de crianças que vivem em domicílios em que nenhum dos moradores tem o ensino fundamental completo (peso: 0,125); 8) Percentual de pessoas de 15 a 24 anos que não estudam, não trabalham e são vulneráveis à pobreza, na população total dessa faixa etária (peso: 0,125).
17	ivs_renda_e_trabalho	IVS Renda e Trabalho	Índice da dimensão Renda e Trabalho, é um dos 3 índices que compõem o IVS. Obtido através da média ponderada de índices normalizados construídos a partir dos indicadores que compõem esta dimensão, a saber: 1) Proporção de vulneráveis à pobreza (peso: 0,200); 2) Taxa de desocupação da população de 18 anos ou mais de idade (peso: 0,200); 3) Percentual de pessoas de 18 anos ou mais sem fundamental completo e em ocupação informal (peso: 0,200); 4) Percentual de pessoas em domicílios vulneráveis à pobreza e dependentes de idosos (peso: 0,200); 5) Taxa de atividade das pessoas de 10 a 14 anos de idade (peso: 0,200).

18	t sem agua esg oto	% de pessoas em domicílios com abastecimento de água e esgotamento sanitário inadequados	Razão entre o número de pessoas que vivem em domicílios cujo abastecimento de água não provém de rede geral e cujo esgotamento sanitário não é realizado por rede coletora de esgoto ou fossa séptica, e a população total residente em domicílios particulares permanentes, multiplicada por 100. São considerados apenas os domicílios particulares permanentes.
19	t sem lixo	% da população que vive em domicílios urbanos sem o serviço de coleta de lixo	Razão entre a população que vive em domicílios sem coleta de lixo e a população total residente em domicílios particulares permanentes, multiplicada por 100. Estão incluídas as situações em que a coleta de lixo é realizada diretamente por empresa pública ou privada, ou o lixo é depositado em caçamba, tanque ou depósito fora do domicílio, para posterior coleta pela prestadora do serviço. São considerados apenas os domicílios particulares permanentes, localizados em área urbana.
20	t vulner mais1h	% de pessoas que vivem em domicílios com renda per capita inferior a meio salário mínimo (de 2010) e que gastam mais de uma hora até o trabalho	Razão entre o número de pessoas ocupadas, de 10 anos ou mais de idade, que vivem em domicílios com renda per capita inferior a meio salário mínimo, de agosto de 2010, e que gastam mais de uma hora em deslocamento até o local de trabalho, e o total de pessoas ocupadas nessa faixa etária que vivem em domicílios com renda per capita inferior a meio salário mínimo, de agosto de 2010, e que retornam diariamente ao trabalho, multiplicado por 100.
21	t mort1	Mortalidade até 1 ano de idade	Número de crianças que não deverão sobreviver ao primeiro ano de vida, em cada mil crianças nascidas vivas.
22	t_c0a5_fora	% de crianças de 0 a 5 anos que não frequentam a escola	Razão entre o número de crianças de 0 a 5 anos de idade que não frequentam creche ou escola, e o total de crianças nesta faixa etária (multiplicada por 100).
23	t_c6a14_fora	% de pessoas de 6 a 14 anos que não frequentam a escola	Razão entre o número de pessoas de 6 a 14 anos que não frequentam a escola, e o total de pessoas nesta faixa etária (multiplicada por 100).
24	t_m10a17_filho	% de mulheres de 10 a 17 anos que tiveram filhos	Razão entre o número de mulheres de 10 a 17 anos de idade que tiveram filhos, e o total de mulheres nesta faixa etária (multiplicada por 100).
25	t_mchefe_fundin _fmenor	% de mães chefes de família, sem fundamental completo e com filho menor de 15 anos de idade	Razão entre o número de mulheres que são responsáveis pelo domicílio, que não têm o ensino fundamental completo e têm pelo menos um filho de idade inferior a 15 anos morando no domicílio, e o número total de mulheres chefes de família (multiplicada por 100). São considerados apenas os domicílios particulares permanentes.

26	t_analf_15m	Taxa de analfabetismo da população de 15 anos ou mais de idade	Razão entre a população de 15 anos ou mais de idade que não sabe ler nem escrever um bilhete simples, e o total de pessoas nesta faixa etária (multiplicada por 100).
27	t_cdom_fundin	% de crianças que vivem em domicílios em que nenhum dos moradores tem o ensino fundamental completo	Razão entre o número de crianças de até 14 anos que vivem em domicílios em que nenhum dos moradores tem o ensino fundamental completo, e a população total nesta faixa etária residente em domicílios particulares permanentes (multiplicada por 100).
28	t_p15a24_nada	% de pessoas de 15 a 24 anos que não estudam, não trabalham e possuem renda domiciliar per capita igual ou inferior a meio salário mínimo (de 2010)	Razão entre as pessoas de 15 a 24 anos que não estudam, não trabalham e com renda per capita inferior a meio salário mínimo, de agosto de 2010, e a população total nesta faixa etária (multiplicada por 100). São considerados apenas os domicílios particulares permanentes.
29	t_vulner	Proporção de pessoas com renda domiciliar per capita igual ou inferior a meio salário mínimo (de 2010)	Proporção de pessoas com renda domiciliar per capita igual ou inferior a meio salário mínimo (2010)
30	t_desocup18m	Taxa de desocupação da população de 18 anos ou mais de idade	Percentual da população economicamente ativa (PEA) nessa faixa etária que estava desocupada, ou seja, que não estava ocupada na semana anterior à data do censo, mas havia procurado trabalho ao longo do mês anterior à data dessa pesquisa.
31	t_p18m_fundin_informal	% de pessoas de 18 anos ou mais sem fundamental completo e em ocupação informal	Razão entre as pessoas de 18 anos ou mais sem fundamental completo, em ocupação informal, e a população total nesta faixa etária, multiplicada por 100. Ocupção informal implica que trabalham, mas não são: empregados com carteira assinada, militares do exército, da marinha, da aeronáutica, da polícia militar ou do corpo de bombeiros, empregados pelo regime jurídico dos funcionários públicos ou empregadores e trabalhadores por conta própria com contribuição a instituto de previdência oficial.
32	t_vulner_depende_idosos	% de pessoas em domicílios com renda per capita inferior a meio salário mínimo (de 2010) e dependentes de	Razão entre as pessoas que vivem em domicílios vulneráveis à pobreza (com renda per capita igual ou inferior a meio salário mínimo de agosto de 2010) e nos quais a renda de moradores com 65 anos ou mais de idade (idosos) corresponde a mais da metade do total da renda domiciliar, e a população total residente em

		idosos	domicílios particulares permanentes (multiplicada por 100).
33	t_atividade10a14	Taxa de atividade das pessoas de 10 a 14 anos de idade	Razão entre as pessoas de 10 a 14 anos de idade que eram economicamente ativas, ou seja, que estavam ocupadas ou desocupadas na semana de referência do censo e o total de pessoas nesta faixa etária (multiplicada por 100). Considera-se desocupada a pessoa que, não estando ocupada na semana de referência, havia procurado trabalho no mês anterior a essa pesquisa.
34	idhm	IDHM	Índice de Desenvolvimento Humano Municipal. Média geométrica dos índices das dimensões Renda, Educação e Longevidade, com pesos iguais.
35	idhm_long	IDHM Longevidade	Índice da dimensão Longevidade que é um dos 3 componentes do IDHM. É obtido a partir do indicador *esperança de vida ao nascer,* através da fórmula: [(valor observado do indicador) - (valor mínimo)] / [(valor máximo) - (valor mínimo)], onde os valores mínimo e máximo são 25 e 85 anos, respectivamente.
36	idhm_educ	IDHM Educação	Índice sintético da dimensão Educação que é um dos 3 componentes do IDHM. É obtido através da média geométrica do subíndice de frequência de crianças e jovens à escola, com peso de 2/3, e do subíndice de escolaridade da população adulta, com peso de 1/3.
37	idhm_renda	IDHM Renda	Índice da dimensão Renda que é um dos 3 componentes do IDHM. É obtido a partir do indicador Renda per capita, através da fórmula: [ln (valor observado do indicador) - ln (valor mínimo)] / [ln (valor máximo) - ln (valor mínimo)], onde os valores mínimo e máximo são R$ 8,00 e R$ 4.033,00 (a preços de agosto de 2010).
38	espvida	Esperança de vida ao nascer	Número médio de anos que as pessoas deverão viver a partir do nascimento, se permanecerem constantes ao longo da vida o nível e o padrão de mortalidade por idade prevalecentes no ano do Censo.
39	idhm_educ_sub_esc	Subíndice de escolaridade - IDHM Educação	Subíndice selecionado para compor o IDHM Educação, representando o nível de escolaridade da população adulta. É obtido pelo indicador *% de jovens e adultos com 18 anos ou mais com o fundamental completo.*

40	idhm educ sub freq	Subíndice de frequência escolar - IDHM Educação	Subíndice selecionado para compor o IDHM Educação, representando a frequência de crianças e jovens à escola em séries adequadas à sua idade. É obtido através da média aritmética simples de 4 indicadores: *% de crianças de 5 a 6 anos na escola*, *% de crianças de 11 a 13 anos no 2º ciclo do fundamental*, *% de jovens de 15 a 17 anos com o fundamental completo* e *% de jovens de 18 a 20 anos com o médio completo.*
41	t_pop18m_fundc	% de 18 anos ou mais com fundamental completo	Razão entre a população de 18 anos ou mais de idade que concluiu o ensino fundamental, em quaisquer de suas modalidades (regular seriado, não seriado, EJA ou supletivo) e o total de pessoas nesta faixa etária multiplicado por 100.
42	t pop5a6 escola	% de 5 a 6 anos na escola	Razão entre a população de 5 a 6 anos de idade que estava frequentando a escola, em qualquer nível ou série e a população total nesta faixa etária multiplicado por 100.
43	t_pop11a13_ffun	% de 11 a 13 anos nos anos finais do fundamental ou com fundamental completo	Razão entre a população de 11 a 13 anos de idade que frequenta os quatro anos finais do fundamental (do 6º ao 9º ano desse nível de ensino) ou que já concluiu o fundamental e a população total nesta faixa etária multiplicado por 100.
44	t pop15a17 fundc	% de 15 a 17 anos com fundamental completo	Razão entre a população de 15 a 17 anos de idade que concluiu o ensino fundamental, em quaisquer de suas modalidades (regular seriado, não seriado, EJA ou supletivo) e o total de pessoas nesta faixa etária multiplicado por 100.
45	t_pop18a20_medioc	% de 18 a 20 anos com médio completo	Razão entre a população de 18 a 20 anos de idade que já concluiu o ensino médio em quaisquer de suas modalidades (regular seriado, não seriado, EJA ou supletivo) e o total de pessoas nesta faixa etária multiplicado por 100. As pessoas de 18 a 20 anos frequentando a 4ª série do ensino médio foram consideradas como já tendo concluído esse nível de ensino.
46	renda per capita	Renda per capita	Razão entre o somatório da renda de todos os indivíduos residentes em domicílios particulares permanentes e o número total desses indivíduos. Valores em reais de 01/agosto de 2010.
47	prosp soc	Prosperidade Social	Nível de prosperidade social da territorialidade, gerada através do cruzamento entre sua faixa do IDHM e do IVS.
48	populacao	População total	População total
49	t fmor5	Mortalidade até 5 anos de idade	Probabilidade de morrer entre o nascimento e a idade exata de 5 anos, por 1000 crianças nascidas vivas.

50	t_razdep	Razão de dependência	Razão de dependência é medida pela razão entre o número de pessoas com 14 anos ou menos e de 65 anos ou mais de idade (população dependente) e o número de pessoas com idade de 15 a 64 anos (população potencialmente ativa) multiplicado por 100.
51	t_fectot	Taxa de fecundidade total	Número médio de filhos que uma mulher deverá ter ao terminar o período reprodutivo (15 a 49 anos de idade).
52	t_env	Taxa de envelhecimento	Razão entre a população de 65 anos ou mais de idade e a população total multiplicado por 100.
53	vulner15a24	População vulnerável de 15 a 24 anos	População de 15 a 24 anos de idade que reside em domicílios com renda per capita igual ou inferior a meio salário mínimo (de agosto de 2010)
54	mchefe_fmenor	Mulheres chefes de família e com filhos menores de 15 anos	População de mulheres que são chefes de família e que possuem pelo menos um filho menor de 15 anos de idade residindo no domicílio.
55	vulner_dia	População ocupada vulnerável à pobreza que retorna diariamente do trabalho	População ocupada vulnerável à pobreza (com renda per capita igual ou inferior a meio salário mínimo de agosto de 2010) e que retornam diariamente do trabalho ao domicílio. São considerados apenas domicílios particulares permanentes.
56	dom_vulner_idoso	População em domicílios vulneráveis e com idoso	População residente em domicílios vulneráveis à pobreza (com renda per capita igual ou inferior à meio salário mínimo de agosto de 2010) em que pelo menos um dos moradores possui idade igual ou superior à 65 anos (idoso). São cosiderados apenas domicílios particulares permanentes.
57	pop0a1	População de até 1 ano	População nessa faixa etária
58	pop1a3	População de 1 a 3 anos	População nessa faixa etária
59	pop4	População de 4 anos	População nessa faixa etária
60	pop5	População de 5 anos	População nessa faixa etária
61	pop6	População de 6 anos	População nessa faixa etária
62	pop6a10	População de 6 a 10 anos	População nessa faixa etária
63	pop6a17	População de 6 a 17 anos	População nessa faixa etária
64	pop11a13	População de 11 a 13 anos	População nessa faixa etária
65	pop11a14	População de 11 a 14 anos	População nessa faixa etária
66	pop12a14	População de 12 a 14 anos	População nessa faixa etária
67	pop15m	População de 15 anos ou mais	População nessa faixa etária

68	pop15a17	População de 15 a 17 anos	População nessa faixa etária
69	pop15a24	População de 15 a 24 anos	População nessa faixa etária
70	pop16a18	População de 16 a 18 anos	População nessa faixa etária
71	pop18m	População de 18 anos ou mais	População nessa faixa etária
72	pop18a20	População de 18 a 20 anos	População nessa faixa etária
73	pop18a24	População de 18 a 24 anos	População nessa faixa etária
74	pop19a21	População de 19 a 21 anos	População nessa faixa etária
75	pop25m	População de 25 anos ou mais	População nessa faixa etária
76	pop65m	População de 65 anos ou mais	População nessa faixa etária
77	pea10m	PEA - 10 anos ou mais	População economicamente ativa. Corresponde ao número de pessoas nessa faixa etária que, na semana de referência do Censo, encontravam-se ocupadas no mercado de trabalho ou que, encontrando-se desocupadas, tinham procurado trabalho no mês anterior à data da pesquisa.
78	pea10a14	PEA - 10 a 14 anos	População economicamente ativa. Corresponde ao número de pessoas nessa faixa etária que, na semana de referência do Censo, encontravam-se ocupadas no mercado de trabalho ou que, encontrando-se desocupadas, tinham procurado trabalho no mês anterior à data da pesquisa.
79	pea15a17	PEA - 15 a 17 anos	População economicamente ativa. Corresponde ao número de pessoas nessa faixa etária que, na semana de referência do Censo, encontravam-se ocupadas no mercado de trabalho ou que, encontrando-se desocupadas, tinham procurado trabalho no mês anterior à data da pesquisa.
80	pea18m	PEA - 18 anos ou mais	População economicamente ativa. Corresponde ao número de pessoas nessa faixa etária que, na semana de referência do Censo, encontravam-se ocupadas no mercado de trabalho ou que, encontrando-se desocupadas, tinham procurado trabalho no mês anterior à data da pesquisa.
81	t_eletrica	% da população em domicílios com energia elétrica	Razão entre a população que vive em domicílios particulares permanentes com iluminação elétrica e a população total residente em domicílios particulares permanentes multiplicado por 100. Considera-se iluminação proveniente ou não de uma rede geral, com

			ou sem medidor.
82	t densidadem2	% da população em domicílios com densidade > 2	Razão entre a população que vive em domicílios particulares permanentes com densidade superior a 2 e a população total residente em domicílios particulares permanentes multiplicado por 100. A densidade do domicílio é dada pela razão entre o total de moradores do domicílio e o número total de cômodos usados como dormitório.
83	rdpc_def_vulner	Renda per capita dos vulneráveis à pobreza	Média da renda domiciliar per capita das pessoas com renda domiciliar per capita igual ou inferior a R$ 255,00 mensais, a preços de agosto de 2010. O universo de indivíduos é limitado àqueles que vivem em domicílios particulares permanentes.
84	t analf 18m	Taxa de analfabetismo - 18 anos ou mais	Razão entre a população de 18 anos ou mais de idade que não sabe ler nem escrever um bilhete simples e o total de pessoas nessa faixa etária multiplicado por 100.
85	t analf 25m	Taxa de analfabetismo - 25 anos ou mais	Razão entre a população de 25 anos ou mais de idade que não sabe ler nem escrever um bilhete simples e o total de pessoas nesta faixa etária multiplicado por 100.
86	t renda trab	% da renda proveniente de rendimentos do trabalho	Participação percentual das rendas provenientes do trabalho (principal e outros) na renda total, considerando-se apenas as pessoas que vivem em domicílios particulares permanentes.
87	i_gini	Índice de Gini	Mede o grau de desigualdade existente na distribuição de indivíduos segundo a renda domiciliar per capita. Seu valor varia de 0, quando não há desigualdade (a renda domiciliar per capita de todos os indivíduos tem o mesmo valor), a 1, quando a desigualdade é máxima (apenas um indivíduo detém toda a renda).O universo de indivíduos é limitado àqueles que vivem em domicílios particulares permanentes.
88	t carteira 18m	% de empregados com carteira - 18 anos ou mais	Razão entre o número de empregados de 18 anos ou mais de idade com carteira de trabalho assinada e o número total de pessoas ocupadas nessa faixa etária multiplicado por 100.
89	t scarteira 18m	% de empregados sem carteira - 18 anos ou mais	Razão entre o número de empregados de 18 anos ou mais de idade sem carteira de trabalho assinada e o número total de pessoas ocupadas nessa faixa etária multiplicado por 100.

90	t_setorpublico_18m	% de trabalhadores do setor público - 18 anos ou mais	Razão entre o número de trabalhadores do setor público de 18 anos ou mais de idade e o número total de pessoas ocupadas nessa faixa etária multiplicado por 100. Os trabalhadores do setor público incluem os empregados pelo regime jurídico dos funcionários públicos e os militares do exército, marinha, aeronáutica, polícia militar ou corpo de bombeiros.
91	t_contapropria_18m	% de trabalhadores por conta própria - 18 anos ou mais	Razão entre o número de trabalhadores por conta própria de 18 anos ou mais de idade e o número total de pessoas ocupadas nessa faixa etária multiplicado por 100.
92	t_empregador_18m	% de empregadores - 18 anos ou mais	Razão entre o número de empregadores de 18 anos ou mais de idade e o número total de pessoas ocupadas nessa faixa etária multiplicado por 100.
93	t_formal_18m	Grau de formalização dos ocupados - 18 anos ou mais	Razão entre o número de pessoas de 18 anos ou mais de idade formalmente ocupadas e o número total de pessoas ocupadas nessa faixa etária multiplicado por 100. Foram considerados como formalmente ocupados os empregados com carteira de trabalho assinada, os militares do exército, da marinha, da aeronáutica, da polícia militar ou do corpo de bombeiros, os empregados pelo regime jurídico dos funcionários públicos, assim como os empregadores e trabalhadores por conta própria que eram contribuintes de instituto de previdência oficial.
94	t_fundc_ocup18m	% dos ocupados com fundamental completo - 18 anos ou mais	Razão entre o número de pessoas de 18 anos ou mais de idade ocupadas que já concluíram o ensino fundamental (regular seriado, regular não seriado, EJA ou supletivo) e o número total de pessoas ocupadas nessa faixa etária multiplicado por 100.
95	t_medioc_ocup18m	% dos ocupados com médio completo - 18 anos ou mais	Razão entre o número de pessoas de 18 anos ou mais de idade ocupadas que já concluíram o ensino médio (regular seriado, regular não seriado, EJA ou supletivo) e o número total de pessoas ocupadas nessa faixa etária multiplicado por 100. Foram consideradas como já tendo concluído o médio aquelas pessoas que frequentavam a 4ª série desse nível de ensino.
96	t_superc_ocup18m	% dos ocupados com superior completo - 18 anos ou mais	Razão entre o número de pessoas de 18 anos ou mais de idade ocupadas e que já concluíram a graduação do ensino superior e o número total de pessoas ocupadas nessa faixa etária multiplicado por 100.
97	t_renda_todos_trabalhos	Rendimento médio dos ocupados - 18 anos ou mais	Média dos rendimentos de todos os trabalhos das pessoas ocupadas de 18 anos ou mais de idade. Valores em reais de agosto de 2010.
98	t_nremunerado_18m	% dos ocupados sem rendimento - 18 anos ou mais	Razão entre o número de pessoas de 18 anos ou mais de idade ocupadas e sem rendimento do trabalho e o número total de pessoas ocupadas nessa faixa etária

			multiplicado por 100.
99	Label_sexo	Sexo	Desagregação das informações por sexo.
100	Label_cor	Cor	Desagregação das informações por cor.
101	Label_sit_dom	Situação de Domicílio	Desagregação das informações por situação de domicílio.

Quadro 8.2: Atributos da Atividade Econômica dos Municípios e Regiões Metropolitanas

Fonte: https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html
Arquivos: Pib_Municipios

Nº	Variável	Tipo de Dado
1	Ano	numérico
2	Código da Grande Região	numérico
3	Nome da Grande Região	caracter
4	Código da Unidade da Federação	numérico
5	Sigla da Unidade da Federação	caracter
6	Nome da Unidade da Federação	caracter
7	Código do Município	numérico
8	Nome do Município	caracter
9	Região Metropolitana	caracter
10	Código da Mesorregião	numérico
11	Nome da Mesorregião	caracter
12	Código da Microrregião	numérico
13	Nome da Microrregião	caracter
14	Código da Região Geográfica Imediata	numérico
15	Nome da Região Geográfica Imediata	caracter
16	Município da Região Geográfica Imediata	caracter
17	Código da Região Geográfica Intermediária	numérico
18	Nome da Região Geográfica Intermediária	caracter
19	Município da Região Geográfica Intermediária	caracter
20	Código Concentração Urbana	numérico
21	Nome Concentração Urbana	caracter
22	Tipo Concentração Urbana	caracter
23	Código Arranjo Populacional	numérico
24	Nome Arranjo Populacional	caracter
25	Hierarquia Urbana	caracter
26	Hierarquia Urbana (principais categorias)	caracter
27	Código da Região Rural	numérico
28	Nome da Região Rural	caracter
29	Região rural (segundo classificação do núcleo)	caracter
30	Amazônia Legal	caracter
31	Semiárido	caracter
32	Cidade-Região de São Paulo	caracter
33	Valor adicionado bruto da Agropecuária, a preços correntes	numérico
34	Valor adicionado bruto da Indústria, a preços correntes	numérico
35	Valor adicionado bruto dos Serviços, a preços correntes – exceto Administração, defesa, educação e saúde públicas e seguridade social	numérico
36	Valor adicionado bruto da Administração, defesa, educação e saúde públicas e seguridade social, a preços correntes	numérico
37	Valor adicionado bruto total, a preços correntes	numérico

38	Impostos, líquidos de subsídios, sobre produtos, a preços correntes	numérico
39	Produto Interno Bruto, a preços correntes	numérico
40	Produto Interno Bruto *per capita*, a preços correntes	numérico
41	Atividade econômica com maior valor adicionado bruto	caracter
42	Atividade econômica com segundo maior valor adicionado bruto	caracter
43	Atividade econômica com terceiro maior valor adicionado bruto	caracter

Quadro 8.3: Atributos do Índice Firjan de Gestão Fiscal de 2013 a 2018

Fonte: https://www.firjan.com.br/ifgf/downloads/download-ifgf-indice-firjan-de-gestao-fiscal.htm
Arquivo: IFGF – Série Histórica

Nº	Atributo	Descrição
1	Código	Código utilizado pelo IBGE para identificação do município.
2	UF	Código da Unidade da Federação
3	Município	Nome do Município
4	Ranking Estadual 2013	Posição do município no Estado
5	Ranking IFGF 2013	Posição do município no Brasil
6	IFGF Geral 2013	Índice Firjan de Gestão Fiscal
7	IFGF Autonomia 2013	(a)
8	IFGF Gastos com Pessoal 2013	(b)
9	IFGF Liquidez 2013	(c)
10	IFGF Investimentos 2013	(d)
11	Ranking Estadual 2014	Posição do município no Estado
12	Ranking IFGF 2014	Posição do município no Brasil
13	IFGF Geral 2014	Índice Firjan de Gestão Fiscal
14	IFGF Autonomia 2014	(a)
15	IFGF Gastos com Pessoal 2014	(b)
16	IFGF Liquidez 2014	(c)
17	IFGF Investimentos 2014	(d)
18	Ranking Estadual 2015	Posição do município no Estado
19	Ranking IFGF 2015	Posição do município no Brasil
20	IFGF Geral 2015	Índice Firjan de Gestão Fiscal
21	IFGF Autonomia 2015	(a)
22	IFGF Gastos com Pessoal 2015	(b)
23	IFGF Liquidez 2015	(c)
24	IFGF Investimentos 2015	(d)
25	Ranking Estadual 2016	Posição do município no Estado
26	Ranking IFGF 2016	Posição do município no Brasil
27	IFGF Geral 2016	Índice Firjan de Gestão Fiscal
28	IFGF Autonomia 2016	(a)
29	IFGF Gastos com Pessoal 2016	(b)
30	IFGF Liquidez 2016	(c)
31	IFGF Investimentos 2016	(d)
32	Ranking Estadual 2017	Posição do município no Estado
33	Ranking IFGF 2017	Posição do município no Brasil
34	IFGF Geral 2017	Índice Firjan de Gestão Fiscal
35	IFGF Autonomia 2017	(a)
36	IFGF Gastos com Pessoal	(b)

	2017	
37	IFGF Liquidez 2017	(c)
38	IFGF Investimentos 2017	(d)
39	Ranking Estadual 2018	Posição do município no Estado
40	Ranking IFGF 2018	Posição do município no Brasil
41	IFGF Geral 2018	Índice Firjan de Gestão Fiscal
42	IFGF Autonomia 2018	(a)
43	IFGF Gastos com Pessoal 2018	(b)
44	IFGF Liquidez 2018	(c)
44	IFGF Investimentos 2018	(d)

Observação:

(a)	IFGF Autonomia	Evidencia um dos pontos mais críticos para a gestão fiscal eficiente das prefeituras: a baixa capacidade de se sustentarem.
(b)	IFGF Gastos com Pessoal	A despesa com pessoal é o principal item da despesa do setor público, no caso dos municípios representam metade da Receita Corrente Líquida (RCL), em média.
(c)	IFGF Liquidez	Verifica se os recursos financeiros são suficientes para fazer frente às despesas que foram postergadas para o ano seguinte.
(d)	IFGF Investimentos	Mede a parcela dos investimentos nos orçamentos municipais.

Quadro 8.4: Histórico dos regimes previdenciários dos entes públicos

Fonte: https://www.gov.br/previdencia/pt-br/assuntos/previdencia-no-servico-publico/estatísticas-e-informacoes-dos-rpps-1/arquivos/2020_set/

Arquivo: 2-historico-dos-regimes-previdenciario_atualizacao_de_ago_set_2020_extracao_em_2020-10-06t18_14_22.xlsx

Nº	Atributo	Descrição
1	CNPJ	CNPJ do ente público
2	UF	UF do ente público
3	ENTE	Nome do ente público
4	REGIME	Regime Previdenciário: RPPS, RPPS em extinção, RGPS
5	DATA DE INICIO DO REGIME	Data de início do regime previdenciário adotado
6	DATA DE FIM DO REGIME	Data de fim do regime previdenciário adotado
7	Tipo de Documento	Instrução normativa relacionada ao regime
8	Número (sem Ano)	Número da instrução normativa
9	Data do Documento	Data da instrução normativa
10	Data da Publicação	Data da publicação da instrução normativa
11	Data de Início da Vigência	Data de início da vigência do regime previdenciário
12	Ementa	Detalhamento da instrução normativa

Quadro 8.5: Atributos dos Municípios – Perspectiva complementar e Target

Fonte: Instituição Financeira

Nº	Atributo	Descrição
1	codigo	Código utilizado pelo IBGE para identificação do município.
2	ano	Ano base
3	beneficiarios_plan_saude_priv	Beneficiários de plano de saúde privado
4	capitacao_dep_av_ap_SFN	Captação (depósito a prazo + poupança) no SFN
5	credito livre SFN	Crédito livre no SFN
6	massa_salarial	Massa salarial do setor formal deflacionada pelo IPCA
7	estoque empregos	Estoque de empregos do setor formal, ponderado pela participação dos cinco setores em cada município
8	conexao_banda_larga	Conexões banda larga por mil habitantes
9	diversidade economica	Subsetores econômicos informados
10	qtd_empresas	Quantidade de empresas com mais de cinco funcionários por mil habitantes
11	qtd leitos hospitalares	Número de leitos hospitalares (complementares + internação)
12	qtd_veiculos_leves	Veículos leves por mil habitantes
13	qtd veiculos pesados	Veículos pesados por mil habitantes
14	crecimento_populacional	Estimativa populacional divulgada pelo IBGE
15	quali prof ensino medio	Qualificação profissional percentual de trabalhadores formais com ensino médio completo
16	desc_consolidado	Variável *target* "não *default*" (0) e "*default*" (1)

Fontes: IBGE, Relação Anual de Informações Sociais (RAIS), Datasus, Anatel, Denatran, Bacen (Estban).