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

Modelos Preditivos para Avaliação de Risco de Corrupção de Servidores Públicos Federais

Ricardo Silva Carvalho

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

Orientador
Prof. Dr. Rommel Novaes Carvalho

Coorientador
Prof. Dr. Donald Matthew Pianto

Brasília
2015

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

Silva Carvalho, Ricardo
SR488m Modelos Preditivos para Avaliação de Risco de Corrupção de Servidores Públicos Federais / Ricardo Silva Carvalho; orientador Rommel Novaes Carvalho; co-orientador Donald Matthew Pianto. -- Brasília, 2015.
118 p.

Dissertação (Mestrado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2015.

1. Corrupção. 2. Servidores Públicos Federais. 3. Mineração de Dados. 4. CGU. 5. DIE. I. Novaes Carvalho, Rommel, orient. II. Matthew Pianto, Donald, co-orient. III. Título.

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

Modelos Preditivos para Avaliação de Risco de Corrupção de Servidores Públicos Federais

Ricardo Silva Carvalho

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

Prof. Dr. Rommel Novaes Carvalho (Orientador)
CIC/UnB

Prof. Dr. Pedro Henrique Melo Albuquerque
FACE/UnB

Prof. Dr. Remis Balaniuk
UCB

Prof. Dr. Marcelo Ladeira
Coordenador do Mestrado Profissional em Computação Aplicada

Brasília, 21 de setembro de 2015

Dedicatória

Dedico à minha família.

Agradecimentos

Agradeço à minha família, amigos e professores, aos meus orientador e coorientador, assim como aos colegas de trabalho da Controladoria-Geral da União, especialmente da Diretoria de Pesquisas e Informações Estratégicas.

Resumo

A Controladoria-Geral da União (CGU), por meio da Diretoria de Pesquisas e Informações Estratégicas (DIE), articula ações de produção de informações estratégicas, investigando possíveis irregularidades cometidas por servidores públicos federais. Com quantitativo reduzido de analistas e inúmeras responsabilidades, a DIE necessita de métodos automatizados aplicáveis a grande volume de dados para aferir corruptibilidade de servidores, buscando assim priorização de trabalho e atuação eficaz baseando-se em aspectos de corrupção.

Este trabalho apresenta a aplicação de mineração de dados para gerar modelos preditivos para avaliar risco de corrupção de servidores públicos federais, usando várias bases de dados a que a CGU tem acesso. O processo CRISP-DM é a referência para as fases da mineração de dados.

Inicialmente, o conhecimento dos analistas especialistas em combate à corrupção da DIE é aplicado nas diversas bases de dados disponíveis para extração de informações possivelmente úteis na indicação de corruptibilidade. Os dados levantados são analisados e preparados com o uso de diversas técnicas, como discretização e análise de correlação, para, em seguida, passarem por um processo de seleção. Métodos de regressão – como *Adaptive Lasso* e Regressão *Ridge* – são aplicados objetivando a criação de modelos preditivos.

O modelo de avaliação de risco de corrupção de servidores públicos federais construído ao final do trabalho obteve resultados satisfatórios de aproximadamente 85% de sensibilidade, 81% de precisão e 83% de acurácia – assim como resultados positivos em testes estatísticos corroborando a validade do modelo com nível de confiança de 95%. Em seguida, as regras geradas pelo modelo final foram analisadas, adicionando-se o estudo de casos pontuais, de modo a subsidiar a descoberta do conhecimento obtido com o processo de mineração de dados.

Com a avaliação de risco de corrupção a partir de modelos preditivos, possibilitou-se: uso mais eficiente e eficaz de recursos e pessoal da CGU; um impacto nacional; e fortalecimento do controle prévio. O direcionamento de esforços de auditoria e fiscalização a partir de índices de corruptibilidade sustenta a priorização efetiva de trabalho da CGU. Atinge-se todos os estados do país analisando em larga escala o nível de corrupção dos

mais de um milhão de servidores públicos federais, gerando impacto em âmbito nacional. Finalmente, todos os pólos regionais da CGU são apoiados com uma atuação de controle prévio, fortalecendo o combate à corrupção.

Palavras-chave: Corrupção, Servidores Públicos Federais, Mineração de Dados, Regressão, *Adaptive Lasso*, *Bootstrap*, Discretização, MDLP, CRISP-DM, DIE, CGU

Abstract

The Brazilian Office of the Comptroller General (CGU), through the Department of Research and Strategic Information (DIE), articulates activities of strategic information production, investigating possible irregularities by federal civil servants. With a reduced quantitative of analysts and numerous responsibilities, DIE needs automated methods applicable to large volumes of data to assess civil servants' corruptibility, seeking then work prioritization and effective action based on aspects of corruption.

This work presents a data mining application to generate predictive models to assess risk of corruption of federal civil servants, using various databases that CGU has access to. The CRISP-DM process is the reference to the phases of the data mining.

Initially, the knowledge of DIE's analysts with expertise in fighting corruption is applied in the various databases available to extract potentially useful information in corruptibility indication. The data collected is analyzed and prepared using various techniques, such as correlation analysis and discretization, to then pass through a selection process. Regression methods – like Adaptive Lasso and Ridge Regression – are applied towards the creation of predictive models.

The model to assess risk of corruption of civil servants built at the end of the work obtained satisfactory results of approximately 85% sensitivity, 81% precision and 83% accuracy – as well as positive results in statistical tests confirming the relevance of the model with a confidence level of 95%. Then, the rules generated by the final model were analyzed, aside with the study of individual cases, in order to support the knowledge discovery through the data mining process.

The assessment of risk of corruption with predictive models allows: more efficient and effective use of CGU's resources; a national impact; and strengthening of previous control. The targeting of audit and control efforts from corruptibility indicators sustains effective prioritization of the work of CGU. Every state in the country is reached by analyzing in large scale the level of corruption of the more than one million federal civil servants, generating impact nationwide. Finally, all regional centers of CGU are backed with prior control activities, strengthening the fight against corruption.

Keywords: Corruption, Federal Civil Servant, Data Mining, Regression, Adaptive Lasso, Bootstrap, Discretization, MDLP, CRISP-DM, DIE, CGU

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

É sabido que, atualmente, o tema corrupção encontra-se consolidado na agenda das preocupações brasileiras [40], sendo fundamentalmente necessário seu combate ostensivo e eficiente. A corrupção pública pode ser definida – utilizando como amparo a Lei nº 8.429, de 2 de junho de 1992[1] – como o ato de improbidade ou valimento indevido de cargo público que importe enriquecimento ilícito, cause lesão ao erário ou atente contra os princípios da Administração Pública.

Já a avaliação de risco de corrupção consiste na identificação de fragilidades que possibilitem a ocorrência de atos de corrupção [25], tornando possível, após uma ponderação dos indicadores descobertos, estimar o risco para um dado cenário. A partir dessa avaliação pode-se orientar, de forma objetiva, o processo de fiscalização para áreas que tenham indícios de tentativas ou perpetração de fraudes, aumentando-se assim o potencial de detecção e de apuração desses eventos.

Pode-se notar que os agentes possuem papel crucial na consecução de práticas de corrupção e uma avaliação de risco relevante neste cenário é aquela envolvendo um grau de corruptibilidade de tais agentes – no caso da corrupção do serviço público federal, os agentes são os próprios servidores públicos federais.

Tendo em vista a avaliação de risco de corrupção para servidores públicos federais, é fato que sua elaboração requer cada vez mais o processamento de uma maior quantidade de informações de diferentes fontes, sejam elas provenientes ou não de sistemas informatizados. Somente o Sistema Integrado de Administração Financeira do Governo Federal (SIAFI)[2] registrou em 2003 uma média mensal de 48 milhões de transações financeiras e em 2013 de 80 milhões[3], representando um aumento de 66% em 10 anos. Além disso, dada a enorme gama de variáveis envolvidas em cenários de corrupção, são necessários

[1] Lei nº 8.429, de 2 de junho de 1992: http://www.planalto.gov.br/ccivil_03/leis/l8429.htm

[2] Site principal do SIAFI: http://www.tesouro.fazenda.gov.br/siafi

[3] Consulta de estatísticas do SIAFI: http://consulta.tesouro.fazenda.gov.br/Estatisticas_novosite/index_estatistica_uso_generica.asp?op=2

modelos preditivos robustos para avaliar risco de corrupção, baseados em informações e hipóteses reais. Por exemplo, pode-se considerar informações funcionais das mais diversas – tempo de serviço, salário, função, atividades –, relações de servidores com empresas, irregularidades de diferentes níveis cadastradas, filiação partidária, relacionamento com terceirizados, etc.

Dessa forma, o uso de técnicas de mineração de dados é extremamente útil em diversos aspectos para elaboração de modelos preditivos para avaliação de risco de corrupção. Inicialmente pode haver a eliminação de subconjuntos redundantes das informações disponíveis nos processos de análise, diminuindo consideravelmente a quantidade de dados tratada. É possível também realizar uma avaliação dos dados relevantes de modo a descobrir padrões que gerem um modelo preditivo de risco de corrupção em função das informações de servidores públicos federais disponíveis. Além disso, pode-se utilizar técnicas de inferência de conhecimento que levam em consideração a incerteza inerente aos processos de identificação de risco de corrupção para gerar modelos preditivos utilizando como insumo, além do suporte de diversas técnicas automatizadas de mineração de dados, a experiência de especialistas em combate à corrupção.

Diante do exposto, observando não somente os aspectos tecnológicos envolvidos, mas também as nuances existentes em cenários de corrupção, este trabalho visa ao estudo e à aplicação de técnicas de mineração de dados para a criação de modelos preditivos para avaliação de risco de corrupção de servidores públicos federais com o intuito de subsidiar o combate à corrupção realizado pela Controladoria-Geral da União (CGU). A partir das mais diversas informações disponíveis sobre servidores públicos federais e do apoio de especialistas em combate à corrupção, os modelos preditivos fornecerão diversas medidas de risco de corrupção indicando o grau de corruptibilidade de servidores públicos federais.

1.1 Definição do Problema

A Controladoria-Geral da União (CGU), órgão integrante da estrutura da Presidência da República, tem como uma de suas competências assistir direta e imediatamente o Presidente da República nos assuntos e providências relacionados à prevenção e ao combate à corrupção. Por meio de ações de produção de informações estratégicas, a Diretoria de Pesquisas e Informações Estratégicas (DIE) é a área responsável por investigar possíveis irregularidades cometidas por servidores públicos federais.

A CGU é um órgão com pólos regionais em todos os estados do Brasil e cujas operações de fiscalização e controle exigem muitas vezes grandes e onerosos deslocamentos – para locais distantes ou de difícil acesso. Dessa forma, considerando ainda a quantidade reduzida de analistas lotados na DIE – que produz informações estratégicas para todo o

país – e a grande abrangência do trabalho da CGU advinda de suas responsabilidades, é importante haver uma eficaz priorização do trabalho com atuação efetiva em larga escala, aumentando a probabilidade de encontrar focos de risco de corrupção e evitando despesas ou deslocamentos desnecessários.

Além disso, cada vez mais é visto que a prática de corrupção é fortemente influenciada pelos servidores alocados nos serviço público[4] [5] [6], sendo necessário considerá-los na priorização das atividades de combate à corrupção. Dessa forma, a mensuração de corruptibilidade de servidores públicos federais a partir de métodos automatizados aplicados a grandes massas de dados se mostra de grande valia, proporcionando uma atuação mais ostensiva da DIE com foco nos agentes públicos.

1.2 Justificativa do Tema

Em relação a providências relacionadas à prevenção e ao combate à corrupção, a Controladoria-Geral da União (CGU) atua através de sua Diretoria de Pesquisas e Informações Estratégicas (DIE) por meio de ações de produção de informações estratégicas. Uma das responsabilidades da DIE é supervisionar e acompanhar a evolução patrimonial dos servidores públicos federais, buscando aferir, por exemplo, ocorrência de enriquecimento ilícito. No entanto, atualmente há aproximadamente 1 milhão servidores públicos federais ativos[7], todos sujeitos a investigação. Devido à essa grande quantidade de servidores existentes no serviço público federal, a DIE se limita boa parte do tempo a realizar apenas investigações de envolvidos em denúncias ou grandes operações federais, muitas vezes restringindo sua atuação aos casos deflagrados externamente. Dessa forma, é importante haver uma priorização de atividades com base em riscos de envolvimento em corrupção para que a DIE possa agir de forma mais efetiva e proativa.

É importante perceber ainda que a mensuração de influência do fator humano nas atividades de corrupção é de difícil realização, dadas a alta subjetividade, a grande quantidade de informações de potencial relevância e a incerteza inerentes. Assim, a utilização de métodos consistentes, baseados em técnicas de mineração de dados e apoiados por conhecimento de especialistas em combate à corrupção, atende a este cenário, pois trata

[4]Notícia de dezembro de 2013 sobre a Máfia dos Fiscais do município de São Paulo: http://oglobo.globo.com/brasil/mafia-dos-fiscais-de-sp-teria-outros-dez-agentes-envolvidos-11044605

[5]Notícia de março de 2013 sobre a Operação Navalha: http://stj.jusbrasil.com.br/noticias/100393628/stj-julga-17-acusados-da-operacao-navalha

[6]Notícia de abril de 2008 sobre a Máfia dos Vampiros no Ministério da Saúde: http://politica.estadao.com.br/noticias/geral,mafia-dos-vampiros-tem-10-denunciados,155005

[7]Site de apresentação do Sistema Integrado de Administração de Recursos Humanos (SIAPE) do Governo Federal: http://www.siapenet.gov.br/oque.htm

tais aspectos de difícil resolução de forma objetiva e automatizada, gerando conhecimento a partir de dados advindos da própria Administração Pública.

Além disso, atualmente a DIE apoia esforços de auditoria e fiscalização com a geração de informações estratégicas em diferentes níveis de detalhamento. No entanto, o direcionamento das atividades de controle leva em conta principalmente fatores de ambiente, e cada vez mais é visto em resultados de operações federais de combate à corrupção[8] [9] que tal prática é fortemente influenciada pelo conluio de agentes públicos, sendo de grande valia, portanto, considerá-los na priorização das atividades de combate à corrupção.

Desse modo, a criação de modelos preditivos para análise de risco de corrupção de servidores públicos federais utilizando técnicas automatizadas de mineração de dados sustenta a seleção priorizada de investigados em suspeitas de enriquecimento ilícito a partir de embasamento estatístico, podendo aumentar as chances de investigação de corruptos, em um trabalho feito em larga escala, impossível de ser realizado manualmente em tempo hábil com o atual quantitativo de servidores lotados na DIE. Além disso, torna possível o direcionamento de esforços de auditoria e fiscalização levando em consideração unidades com servidores de alto índice de corruptibilidade, possibilitando uma atuação mais consciente da criticidade de cada ambiente observando seus agentes alocados, otimizando o uso de recursos e pessoal da CGU e aumentando o alcance do combate à corrupção. Este trabalho proporciona, dessa maneira, impacto em âmbito nacional abarcando todos os estados em níveis de corrupção de servidores públicos federais, apoiando todos os pólos regionais da CGU em termos de auditoria e fiscalização, onde considerando a atuação de controle prévio – observando valores contabilizados em operações relacionadas à corrupção[10] – há um enorme potencial de economia para os cofres públicos através das práticas de prevenção de corrupção.

1.3 Objetivos

Este trabalho possui dois grandes objetivos. O primeiro é construir um modelo preditivo para avaliação de risco de corrupção de servidores públicos federais. Para tal, busca-se aplicar técnicas de mineração de dados com base no estado da arte, juntamente com um estudo detalhado do cenário no qual as informações relacionadas à corrupção se inserem,

[8] Notícia de agosto de 2011 sobre a Operação Voucher: http://politica.estadao.com.br/noticias/geral,pf-prende-nr-2-do-ministerio-do-turismo-e-mais-37-por-corrupcao,756070

[9] Notícia de abril de 2013 sobre a Operação Nacional contra a Corrupção comandada pelo Ministério Público: http://g1.globo.com/jornal-da-globo/noticia/2013/04/operacao-nacional-contra-corrupcao-prende-92-pessoas-em-12-estados.html

[10] Artigo da Revista Mundo Estranho, edição 122 de março de 2012, sobre os maiores escândalos de corrupção do Brasil, utilizando ranking em termos de prejuízo dos cofres públicos: http://mundoestranho.abril.com.br/materia/os-maiores-escandalos-de-corrupcao-do-brasil

tornando possível gerar um índice de corruptibilidade para cada servidor público federal a partir de dados de diferentes fontes.

O segundo objetivo é realizar a descoberta de conhecimento no que tange a informações sobre corruptibilidade de servidores públicos federais, buscando levantar novas regras desse domínio. Para tal, as informações de servidores públicos disponíveis – assim como suas relações diretas e indiretas – são analisadas com o apoio de especialistas da DIE no combate à corrupção e técnicas de mineração de dados apropriadas são aplicadas.

Como objetivos implícitos, pode-se citar a seleção de atributos relevantes para cada aspecto de corrupção, avaliação de modelos preditivos gerados a partir dos diversos algoritmos de regressão, assim como a comparação entre os resultados obtidos através dos modelos desenvolvidos e o conhecimento atualmente empregado pela equipe de especialistas da DIE. Tal comparação possui como finalidade obter conhecimentos não vislumbrados previamente, validar as premissas existentes e confirmar o ganho na geração de modelos preditivos.

1.4 Metodologia

O presente trabalho tem sua metodologia resumida nos seguintes tópicos:

- Buscando seguir o modelo de referência CRISP-DM [23], iniciou-se com o levantamento dos trabalhos da CGU – e mais especificamente da DIE – relacionados ao combate à corrupção envolvendo irregularidades de servidores públicos federais;
- Foram registradas as diversas bases de dados disponíveis à DIE, com as principais informações contidas em cada contexto;
- Foi feito um estudo inicial para validar o uso de mineração de dados no contexto de risco de corrupção, utilizando-se apenas dados de filiação partidária, em um trabalho de aprendizagem de máquina usando algoritmos de classificação;
- Os resultados do estudo inicial foram discutidos com especialistas em combate à corrupção e professores da área de mineração de dados, optando-se por alterar a abordagem de classificação para regressão, devido principalmente à necessidade de resultados contínuos de risco de corrupção de servidores;
- Realizou-se um estudo das técnicas de regressão mais aplicadas em trabalhos de pesquisa no âmbito de combate a fraudes, assim como de outras técnicas de mineração de dados úteis ao trabalho – como reamostragem e discretização;

- Com o apoio de especialistas em combate à corrupção da DIE, foi feito um levantamento das informações mais relevantes em termos de corruptibilidade que poderiam ser extraídas das bases de dados;
- Utilizando a ferramenta *SQL Server Management Studio*[11] foram obtidos dados das diversas bases disponíveis na DIE;
- Foram realizadas discussões para validação das informações extraídas e definiu-se um conjunto apropriado de informações relacionadas à corrupção a serem extraídas e estudadas;
- Após a extração do conjunto definido de informações, foram realizadas reuniões semanais orientadas pelos dados obtidos, para maior entendimento do significado das informações e do contexto de cada base de dados, assim como dos aspectos relacionados ao contexto de combate à corrupção;
- De posse de um maior entendimento do domínio e com o aprimoramento do estudo de técnicas de mineração de dados, realizou-se o ajuste de dados utilizando-se a ferramenta R através da IDE RStudio[12];
- A partir dos dados preparados foram criados modelos preditivos para avaliação de risco de corrupção de servidores públicos federais, juntamente com diversas métricas de avaliação, seguindo algoritmos bastante utilizados, como *Lasso* para seleção de variáveis e MDLP (*Minimum Description Length Principle*) para discretização;
- Professores da UnB (Universidade de Brasília) foram consultados quanto aos métodos utilizados, com o intuito de validar a abordagem utilizada; e
- Os modelos gerados e suas regras foram analisados em reuniões com especialistas da DIE para definição de novos ajustes e realização de remodelagens até a geração de um modelo preditivo final consistente.

1.5 Contribuições

O presente trabalho busca gerar modelos preditivos que expressem um índice de corruptibilidade para servidores públicos federais a partir de bases de dados disponíveis no âmbito da CGU, especificamente na DIE. Dessa forma, pretende-se contribuir com o trabalho da CGU no que tange ao combate à corrupção, pois o uso dos índices gerados para toda a Administração Pública Federal apoiará a atuação do controle prévio – em

[11] *SQL Server Management Studio*. Link: https://www.microsoft.com/pt-br/download/details.aspx?id=29062

[12] RStudio – IDE para ferramenta estatística R. Link: http://www.rstudio.com/

investigação pontuais de servidores com alto grau de corruptibilidade para, por exemplo, fins de descoberta de ocorrência de enriquecimento ilícito – assim como possibilitará o direcionamento de trabalhos de auditoria e fiscalização do órgão, permitindo a CGU atuar em unidades baseando-se não somente em aspectos do ambiente, mas também em características dos agentes alocados.

Ainda no âmbito do combate à corrupção, com a visualização e o entendimento objetivo das regras embutidas nos modelos gerados, este trabalho pretende contribuir com o arcabouço técnico do que os especialistas entendem ser aspectos que implicam corrupção. Cada regra advinda dos modelos melhor ajustados para cada aspecto de corruptibilidade poderá tanto confirmar dado conhecimento dos especialistas em combate à corrupção como refutá-lo. Além disso, as regras podem também abordar características nunca antes vislumbradas pelos especialistas como influenciadoras em corruptibilidade, gerando assim novos conhecimentos anteriormente apenas implícitos nas bases de dados da Administração Pública Federal.

1.6 Estrutura do Documento

Este documento está estruturado da seguinte forma: após o Capítulo 1 de introdução, o Capítulo 2 detalha a fundamentação teórica tida como base para todo o trabalho, considerando estudos relacionados tanto a aspectos de criação de modelos preditivos e outras técnicas de mineração de dados – por exemplo, reamostragem e discretização – como a trabalhos correlatos aplicados à detecção de fraudes.

No Capítulo 3 tem-se o delineamento dos experimentos realizados no âmbito deste trabalho, elencando as resoluções apontadas para cada problema encontrado, seguindo as fases do modelo de referência CRISP-DM. Já o Capítulo 4 apresenta e faz uma análise dos resultados obtidos na consecução do trabalho com a solução proposta. Finalmente, o Capítulo 5 traz as conclusões obtidas e os trabalhos futuros vislumbrados.

Capítulo 2

Fundamentação Teórica

Neste capítulo, são citados os principais referenciais teóricos e os trabalhos acadêmicos relevantes relacionados. Inicialmente, após uma breve introdução sobre mineração de dados, o modelo de referência para projetos de mineração de dados CRISP-DM [23] é apresentado. Em seguida, são apresentadas duas técnicas de reamostragem, *Bootstrap* [68] e *Cross-validation* [67], úteis para obter estimativas de métricas de distribuição desconhecida [68].

Logo após, algoritmos de regressão a serem utilizados no presente trabalho são descritos, com foco em métodos com regularização [67], como Regressão *Ridge* [34], *Lasso* [66] e *Adaptive Lasso* [73], para realizar seleção de variáveis [67]. Além disso, frisa-se o uso de regressão logística [36], pois necessita-se de modelos com respostas probabilísticas para avaliação de risco de corrupção.

Com o intuito de possibilitar a modelagem de não linearidades, são introduzidos algoritmos de discretização bastante utilizados atualmente [28]. Possuindo abordagens distintas, serão descritos os três algoritmos utilizados no presente trabalho, a saber: CAIM (*Class-Attribute Interdependence Maximization*) [45], MDLP (*Minimum Description Length Principle*) [37] e ModChi2 (*Modified Chi2*)[64].

Além disso, este capítulo apresenta a indicação de técnicas de avaliação de modelos de regressão [67], buscando descrever testes de hipótese [32] e métricas bastante utilizados no meio acadêmico [69], úteis tanto para análise do ajuste dos modelos gerados quanto para comparação entre modelos válidos. Finalmente, trabalhos correlatos envolvendo mineração de dados e corrupção são expostos para análise comparativa.

2.1 Mineração de Dados

A mineração de dados inicialmente se enquadra na noção de busca de informações úteis embutidas em grandes quantidades de dados [71]. É um campo interdisciplinar

que funde conceitos de áreas relacionadas, como bancos de dados [27], estatística [21], aprendizagem de máquina [67] e reconhecimento de padrões [72]. A mineração de dados pode ser vista como parte de um processo maior de descoberta de conhecimento, que inclui tarefas de pré-processamento – como extração e limpeza de dados, reamostragem, redução de dimensionalidade, construção de variáveis – assim como etapas de pós-processamento – como interpretação de modelos e padrões, geração e confirmação de hipóteses, entre outras [72].

Em geral, considerando as tarefas envolvidas em um projeto de mineração de dados, deve-se levar em conta tanto aspectos técnicos em relação à manipulação de dados e à geração de modelos, como entendimento do domínio ou das regras de negócio do ambiente onde o projeto se insere. Com tal objetivo, será detalhada a seguir uma metodologia amplamente utilizada como modelo de referência para projetos de mineração de dados e que envolve todas as nuances necessárias ao trabalho, chamada CRISP-DM (*Cross Industry Standard Process for Data Mining*) [23].

Nesse contexto, a mineração de dados preditiva [71] pode ser denominada como a busca de critérios que possam ser generalizados de forma a possibilitar avaliações futuras. O objetivo da predição é examinar exemplos com rótulos e encontrar padrões generalizáveis que permitam prever rótulos de novos exemplos. Tendo o exposto em vista, o problema onde os rótulos assumem valores categóricos ou discretos é dito de classificação, enquanto que para rótulos assumindo valores contínuos o problema é denominado regressão [71]. Como o foco do presente trabalho é obter índices de risco de corrupção, com valores contínuos de probabilidade situados de 0 a 1, métodos de regressão são estudados objetivando a geração de modelos de corruptibilidade.

2.1.1 CRISP-DM

O CRISP-DM (*Cross Industry Standard Process for Data Mining*) [23] é um modelo de referência para processos de mineração de dados. A metodologia fornece uma visão geral do ciclo de vida de um projeto de mineração de dados, consistindo em seis fases distintas, a saber: Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implantação.

Como a sequência de fases do CRISP-DM não é rígida, na Figura 2.1 é possível observar as dependências mais importantes e frequentes entre as fases, indicadas pelas setas. Já o círculo externo simboliza a natureza cíclica da própria mineração de dados. A seguir as fases mencionadas serão brevemente descritas.

Entendimento do Negócio

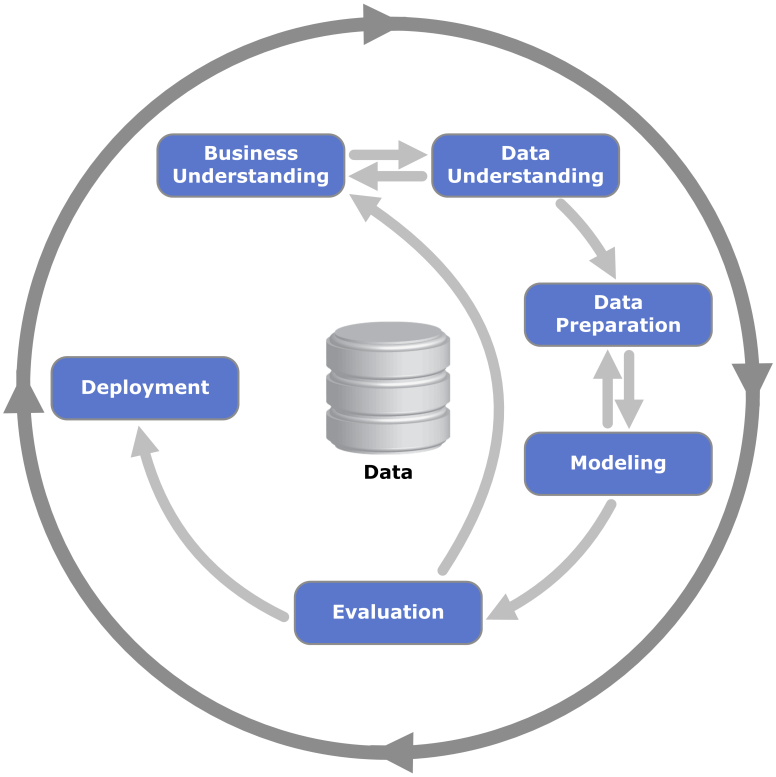

Figura 2.1: Fases do CRISP-DM[1]

A fase inicial denominada Entendimento do Negócio é responsável por determinar os objetivos a serem alcançados. A partir da perspectiva de negócio, esta fase visa compreender o domínio do projeto de mineração de dados, envolvendo a organização e os problemas relacionados. Para tal, as tarefas dessa fase incluem, além da determinação de metas, a avaliação do cenário atual onde o projeto encontra-se inserido, apontando ainda facilitadores e restrições que podem influenciar o sucesso do projeto.

Entendimento dos Dados

O Entendimento dos Dados é a fase responsável pela exploração inicial dos dados, onde os dados são identificados e descritos. Nesta fase, os dados disponíveis são coletados e suas informações de quantidade de registros, campos e formatos são compiladas. A partir deste entendimento busca-se a compreensão das informações contidas nos dados, além da identificação de possíveis problemas de qualidade. Com as informações obtidas desta primeira exploração já é possível apontar atributos relevantes para as fases seguintes.

Preparação dos Dados

[1]Figura das fases do CRISP-DM obtida em: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

A etapa de Preparação dos Dados envolve a construção e formatação dos conjuntos de dados finais. Compreende desde limpeza de dados até integração de dados em formatos distintos. Nesta fase atributos derivados são criados, além da seleção de atributos e outras atividades, como discretização, tratamento de valores faltantes e remoção de atributos correlacionados. Por ser bastante complexa, a preparação dos dados pode ser executada diversas vezes. Com o conjunto final de dados construído, realiza-se aqui também a separação dos dados de modo a criar um conjunto de dados de treinamento, a ser usado na fase modelagem, e um conjunto de dados de teste, não usado no treinamento, mas sim ao final da fase de avaliação como meio de verificação de modelos em dados não observados durante a modelagem.

Modelagem

Na fase de Modelagem são selecionadas as técnicas de mineração de dados para resolver o problema levantado. Nesta fase, algoritmos de aprendizagem de máquina mais adequados a cada cenário são configurados para construir modelos aderentes e compatíveis com os dados preparados. Dessa forma, esta etapa está bastante relacionada com a fase de preparação dos dados.

Avaliação

A Avaliação consiste na análise do modelo construído quanto ao alcance dos objetivos de negócio. Trate-se de métodos para validar o modelo final, onde são revistos os passos executados até sua construção, verificando se alguma tarefa deve ser ajustada. Compreende também a análise das regras geradas pelo modelo final, com sua avaliação segundo critérios de negócio, com o intuito de decidir se o modelo será utilizado ou não. Esta fase pode gerar a necessidade de retornar a qualquer uma das fases anteriores para ajustes das atividades executadas.

Implantação

Finalmente, é na fase de Implantação que os resultados obtidos são colocados em uso ou apresentados aos interessados no projeto, com o intuito de efetivamente atingir os objetivos pretendidos. Nesta fase pode-se tanto implantar um processo na organização em questão que reflita os resultados do processo de mineração de dados, quanto simplesmente gerar relatórios apresentando o conhecimento gerado através do projeto. Na etapa de Implantação planos futuros também podem ser delineados, com vistas a melhoria de resultados.

2.2 Balanceamento de Dados

Atualmente, diversos trabalhos aplicam mineração de dados em cenários contendo conjuntos de dados com desbalanceamento de classes [30], em outras palavras, praticamente todos os registros são de uma classe, enquanto que uma quantidade bem menor de registros é marcada com a outra classe, normalmente a mais importante – esta última pode ser chamada de classe rara.

Em tais casos algoritmos tendem a ficar presos à classe majoritária e ignorar a classe rara, já que os modelos buscam boa performance sob o conjunto completo de registros [30].

Dois métodos usuais para minimizar o desbalanceamento [52] dos dados são: *under-sampling* e *over-sampling*. No primeiro, os dados são balanceados ao se descartar aleatoriamente registros da classe majoritária, enquanto que no segundo, o desbalanceamento é modificado ao serem replicados registros da classe rara.

Em relação a tais métodos, há tanto trabalhos afirmando que não há resposta geral para qual distribuição de classes terá melhor performance [70], dependendo de cada caso, assim como sugerindo manter a distribuição original [30] ou realizar *under-sampling* [39]. No entanto, não há consenso e no geral tanto *under-sampling* quanto *over-sampling* se mostram como métodos efetivos para atacar o problema de desbalanceamento [39] [30].

Vale frisar ainda que há trabalhos relacionando o método de *over-sampling* com o fenômeno de *overfitting* [70] [30], no qual um dado modelo é criado através de um excessivo ajuste ou superajuste aos dados, o que é indesejável pois dificulta sua generalização.

2.3 Reamostragem

Com a atual disponibilização menos custosa de computação rápida, métodos de análise estatística computacionalmente intensivos como a reamostragem tem se tornado cada vez mais comuns [42]. As técnicas de reamostragem são capazes de substituir diversos métodos estatísticos tradicionais ao aplicar reamostragem repetidas vezes nos dados originais e realizar inferências a partir das subamostras [68].

No cenário de mineração de dados, uma aplicação imediata de reamostragem é a execução de algoritmos para descoberta de padrões em diversas subamostras obtidas de um mesmo conjunto de dados, seguindo com uma comparação dos resultados de estatísticas de interesse, observando o quanto diferem. Apesar de tais técnicas poderem se tornar computacionalmente intensivas, por obterem resultados de várias execuções em conjuntos de dados diferentes, com o atual avanço da computação os requisitos computacionais necessários para uso de técnicas de reamostragem geralmente não são proibitivos [68].

Duas técnicas de reamostragem bastante utilizadas atualmente [29] são *Bootstrap* e *Cross-validation*.

2.3.1 *Bootstrap*

Em cenários de cálculo de parâmetros onde as distribuições inerentes são desconhecidas, o *Bootstrap* é útil para fornecer estimativas de variabilidade [68]. A partir de uma mesma amostra inicial são geradas várias subamostras e uma distribuição empírica é calculada a partir das diversas subamostras, não sendo necessário, portanto, assumir nenhuma distribuição previamente[68]. Dessa forma, *Bootstrap* permite aproximar uma distribuição desconhecida pela distribuição empírica dos dados baseando-se apenas em uma amostra única finita. Esta técnica possui diversas variantes, como por exemplo: com ou sem reposição, conhecendo-se (*Bootstrap* paramétrico) ou não (*Bootstrap* não paramétrico) a distribuição dos dados.

Com o uso do *Bootstrap*, pode-se facilmente obter estimativas de parâmetros complexos, juntamente com as estimativas de erro dos mesmos. Por exemplo, no caso de B reamostragens sem reposição, pode-se obter a estimativa de dada estatística de interesse a partir da média dos valores obtidos em cada uma das B subamostras, sendo o desvio-padrão divido pela raiz quadrada do número total de observações o valor da estimativa de erro [68].

É útil perceber ainda que, como busca-se para resultados das estimativas a distribuição dos valores médios, a partir de uma derivação do Teorema do Limite Central, a distribuição dos valores estimados médios deve seguir a distribuição Normal [68]. Logo, tal fato pode ser usado como forma de verificação da adequação dos resultados das estimativas.

2.3.2 *Cross-validation*

A técnica de reamostragem denominada *Cross-validation* [67] é geralmente utilizada em cenários de aprendizagem de máquina, onde são geradas duas subamostras e um modelo é gerado na primeira e validado na segunda. Dessa forma, é bastante utilizado no contexto de seleção de modelos – atividade onde são avaliados diversos modelos diferentes, buscando-se selecionar o melhor segunda alguma métrica [42] –, pois fornece estimativas de métricas de validação menos tendenciosas.

Na prática, o *Cross-validation* divide aleatoriamente um conjunto de dados em k partes mutuamente exclusivas, o que é definido como *k-fold Cross-validation*. Com tal método k modelos são gerados, onde, a cada passo, $k-1$ partes são utilizadas para treinamento e uma única parte é usada na validação. Assim, a estimativa de dada métrica de validação é

tomada com a média dos resultados na parte de validação em cada uma das k execuções. Atualmente o número de partes mais utilizado é tipicamente $k = 10$.

2.4 Métodos de Regressão

Tipicamente em métodos de regressão busca-se prever uma variável aleatória Y – também chamada de variável dependente, saída ou resposta – a partir de uma ou mais variáveis independentes – também chamadas de covariáveis, preditores ou entradas – denotadas como um vetor de entrada $X^T = (X_1, X_2, \ldots, X_p)$ [67].

Anteriormente à existência de computadores, usava-se amplamente no campo da estatística o método de Regressão Linear [69]. Esta técnica supõe que a relação entre Y e X é uma função linear de distribuição Gaussiana e coeficientes α e β, adicionalmente a um erro aleatório ε independente de X e com $E(\varepsilon) = 0$, onde a função ou modelo preditivo gerado é $f(X)$, para p covariáveis, segundo as equações abaixo:

$$Y = f(X) + \varepsilon \tag{2.1}$$

$$f(X) = E[Y|X] = \alpha + \sum_{j=1}^{p} \beta_j X_j \tag{2.2}$$

No entanto, vale observar que em situações práticas a premissa de relação linear poderá não ser válida, o que traz a necessidade da utilização de métodos que possam modelar relacionamentos não lineares [49]. Dessa forma, a seguir serão descritos possíveis métodos amplamente utilizados atualmente para modelar não linearidades [67], aumentando a flexibilidade dos modelos criados mas sem perder a facilidade e interpretabilidade dos modelos lineares. A ideia destes métodos é modificar ou substituir o vetor de covariáveis X por outras variáveis que são transformações de X, para em seguida ajustar normalmente um modelo linear nesse novo espaço de covariáveis derivadas, pois uma vez que as transformações ou funções de X foram definidas, os modelos passam a ser lineares nessas novas covariáveis [67].

É útil frisar ainda que, para casos onde a resposta deve estar situada num intervalo restrito de valores, a regressão linear também não é adequada, pois não limita os valores de saída, supondo apenas uma distribuição Gaussiana da resposta [36]. Assim, como este trabalho busca índices probabilísticos de risco, serão apresentadas soluções que possibilitam aplicar a restrição de que a variável dependente segue uma distribuição de Bernoulli, ou seja, é limitada a valores contínuos de probabilidade situados entre 0 e 1 [21].

Além disso, devido a possibilidade de grande número de covariáveis nos contextos estudados e considerando o objetivo de reduzir o conjunto de covariáveis às mais significativas,

será introduzida a técnica de Regularização [67] – que busca eliminar covariáveis a partir de métodos de encolhimento das estimativas dos coeficientes – seguida de três métodos amplamente utilizados [67], a saber: Regressão *Ridge* [34], *Lasso* [66] e *Adaptive Lasso* [73].

2.4.1 Regressão Local

Em métodos de Regressão Polinomial [21] cria-se um polinômio com novas covariáveis modificadas X_t como potência das originais X – resultando em, por exemplo, $X_{t1} = X_1, X_{t2} = X_1^2, X_{t3} = X_1^3$ – ou como interações das mesmas, resultando em, por exemplo, $X_{t1} = X_1, X_{t2} = X_1X_2, X_{t3} = X_1X_3$. Dessa forma, apesar de descrever não linearidades com a transformação polinomial, após as modificações o modelo é linear nas covariáveis modificadas [67].

Apesar da Regressão Polinomial descrita ser útil para modelar não linearidades, um único polinômio pode não descrever a totalidade dos dados, no caso de, por exemplo, a não linearidade estar concentrada apenas em determinada região dos dados. Tendo isso em vista, os métodos de Regressão Local [24] possibilitam separar regiões dos dados por valores chamados nós, em vez de se utilizar apenas um único polinômio em X para todo o domínio, criando assim diferentes polinômios, um para cada região definida.

LOESS

No método de regressão local LOESS (*Locally Weighted Regression*) [24], a modelagem de não linearidade é ajustada com a realização de regressões lineares locais ponderadas, ou seja, são estimadas funções nas vizinhanças de cada ponto de interesse resolvendo um problema de mínimos quadrados ponderados [67]. Geralmente, os dados mais próximos de cada ponto de interesse possuem um peso maior no modelo, enquanto que os mais distantes possuem um peso menor. Dessa forma, tal método tende a ser robusto e resistente a *outliers*, além de necessitar de poucas observações para uma boa estimativa de como Y depende de X [67]. No ajuste do LOESS, o parâmetro de suavização α representa uma certa porcentagem do conjunto de pontos, definindo assim o número de nós n que serão levados em consideração em cada regressão local [69] – por exemplo, havendo 16 observações e $\alpha = 0.6$, então $n = 9$.

2.4.2 Modelos Lineares Generalizados (MLG)

Os Modelos Lineares Generalizados (MLG) [49] expandem o modelo de regressão linear generalizando a resposta para qualquer distribuição da família exponencial como mostrado

a seguir:

$$g(E[Y|X]) = \alpha + \sum_{j=1}^{p} \beta_j X_j \tag{2.3}$$

Assim, através do uso de uma função de ligação $g(E[Y|X])$, um MLG relaciona o valor esperado da resposta $E[Y|X]$ com as p covariáveis X_j de um modelo de regressão linear e seus coeficientes β_j, além da constante α.

Regressão Logística

No caso de um MLG com a resposta como distribuição de Bernoulli, o modelo resultante é conhecido como Regressão Logística [36]. Tal método relaciona $E[Y|X] = Pr(Y = 1|X)$ com as covariáveis através de uma função de ligação, como por exemplo a *logit*, como mostrado na equação a seguir:

$$logit\left(E[Y|X]\right) = \ln\left(\frac{E[Y|X]}{1 - E[Y|X]}\right) = \alpha + \sum_{j=1}^{p} X_j \beta_j \tag{2.4}$$

Assim, vê-se que $E[Y|X]$ não é mais escrito como uma combinação linear de covariáveis X_j – além de seus coeficientes β_j e constante α –, mas sim $logit(E[Y|X])$ que é colocada como tal combinação, o que transforma a resposta do modelo para a distribuição mencionada – onde ln indica logaritmo neperiano.

Dessa forma, $E[Y|X] = Pr(Y = 1|X)$ representa a probabilidade de $Y = 1$, que pode ser obtida pela expressão:

$$Pr(Y = 1|X) = \frac{e^{\alpha + \sum_{j=1}^{p} X_j \beta_j}}{1 + e^{\alpha + \sum_{j=1}^{p} X_j \beta_j}} \tag{2.5}$$

onde se tem a resposta de um modelo $Pr(Y = 1|X)$ em termos de probabilidade, obtida a partir dos valores das covariáveis X_j combinados com seus coeficientes β_j e a constante α – sendo e o número de Euler, obtido através da equação anterior com aplicação do logaritmo neperiano.

2.4.3 Regressão com Regularização

Nos métodos de regressão, o vetor de coeficientes $\beta = \beta_1, \beta_2, \ldots, \beta_p$ e a constante α geralmente são estimados através de um método para minimização de alguma função [21]. No caso de regressão linear, por exemplo, utiliza-se o método dos mínimos quadrados

para minimizar a soma dos quadrados dos desvios (RSS - *Residual Sum of Squares*) [21], definida como:

$$RSS = \sum_{i=1}^{n} \left(y_i - \alpha - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{2.6}$$

onde y_i é o valor observado de uma resposta i entre n para os valores de p covariáveis x_{ij} com seus coeficientes β_j e constante α.

Vê-se que nessas abordagens usuais, as minimizações incluem todas as covariáveis no modelo final. Assim, no caso onde o número de covariáveis é muito maior que o número de observações a minimização pode não ser possível [67]. Além disso, dado que modelos mais simples são mais fáceis de interpretar, pode ser útil a redução de covariáveis, escolhendo apenas as mais significativas dentro do conjunto completo de covariáveis.

Dessa forma, uma alternativa é utilizar Regularização na função a ser minimizada, de forma que os coeficientes β_j estimados sejam regularizados ou encolhidos em direção a zero [67]. A forma usual de regularizar é adicionar penalidades aos coeficientes na função a ser minimizada. Duas técnicas bastante usadas para regularização [67] são Regressão *Ridge* [34] e *Lasso* [66].

Regressão *Ridge*

Na Regressão *Ridge* [34], utiliza-se a penalidade $L2$ descrita como:

$$Penalidade\ L2 = \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.7}$$

onde λ é o parâmetro de ajuste da penalidade, que controla o impacto da regularização.

Como a penalidade é adicionada à função a ser minimizada, os valores da Equação 2.7 acima deverão também ser minimizados, o que levará ao encolhimento dos coeficientes β_j em direção a zero. É válido observar que valores de λ diferentes levarão a estimativas distintas, logo sua escolha é crítica na Regressão *Ridge*. Para a definição do λ utiliza-se usualmente *Cross-Validation* [67].

Lasso

Apesar da redução dos coeficientes β_j, a Regressão *Ridge* ainda mantém todos as p covariáveis no modelo final. Mesmo aumentando o valor de λ, a penalidade $L2$ definida na Equação 2.7 irá encolher os coeficientes em direção a zero, mas não definirá nenhum deles como exatamente zero. Apesar disso possivelmente não gerar problemas em relação

à acurácia do modelo, ainda dificulta sua interpretação, pois todas as covariáveis são mantidas [67].

O método *Lasso* [66] supera essa desvantagem da Regressão *Ridge*, ao utilizar a penalidade $L1$, definida como:

$$Penalidade\ L1 = \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.8}$$

onde λ é o parâmetro de ajuste da penalidade, que controla o impacto da regularização.

Assim como na Regressão *Ridge*, *Lasso* também encolhe os coeficientes em direção a zero. No entanto, a penalidade $L1$ definida na Equação 2.8 acima força algumas estimativas dos coeficientes β_j serem exatamente iguais a zero quando o parâmetro λ é suficientemente grande. Por exemplo, em dado caso onde a Regressão *Ridge* define um coeficiente de uma covariável como próximo de zero – mas ainda diferente de zero, ou seja, mantendo a covariável respectiva no modelo, apesar do coeficiente pequeno – *Lasso* pode gerar uma estimativa de coeficiente exatamente igual a zero, eliminando a covariável correspondente. Como resultado, *Lasso* realiza uma seleção de variáveis, resultando em um modelo mais facilmente interpretável. Assim como na Regressão *Ridge*, a escolha do valor de λ é crítica, onde usualmente utiliza-se *Cross-Validation* para tal [67].

Apesar de amplamento utilizado, o *Lasso* produz estimativas tendenciosas para coeficientes grandes, em outras palavras, *Lasso* não atende a propriedade de oráculo [73]. Um método com tal propriedade é capaz de estimar o subconjunto das verdadeiras covariáveis com coeficientes exatamente iguais a zero com probabilidade tendendo a 1, ou seja, produz estimativas de coeficientes como se o subconjunto referido já fosse conhecido previamente. Assim, um estimador que atende a propriedade de oráculo realiza uma seleção de covariáveis assintoticamente consistente e eficiente, com estimativas não tendenciosas para grandes coeficientes.

Adaptive Lasso

Para resolver o problema da propriedade de oráculo, o *Adaptive Lasso* [73] foi proposto, com a penalidade definida como:

$$Penalidade\ Adaptive\ Lasso = \lambda \sum_{j=1}^{p} \hat{\omega}_j |\beta_j| \tag{2.9}$$

onde λ é o parâmetro de ajuste da penalidade, que controla o impacto da regularização, e $\hat{\omega}_j\ (j = 1, \ldots, p)$ é o vetor adaptativo de pesos.

O vetor adaptativo de pesos $\hat{\omega}_j$ pode ser definido como:

$$\hat{\omega}_j = \frac{1}{\left(|\hat{\beta}_j^{ini}|\right)^{\gamma}} \tag{2.10}$$

onde γ é uma constante positiva e $\hat{\beta}_j^{ini}$ é uma estimativa inicial consistente dos coeficientes β_j usualmente obtida através da execução prévia da Regressão *Ridge*. Assim como no *Lasso*, λ é obtido através de *Cross-Validation*. Já a constante γ é sugerida ser selecionada entre os valores 0.5, 1 e 2 [73].

O vetor adaptativo de pesos permite ao *Adaptive Lasso* aplicar diferentes níveis de encolhimento ou regularização a diferentes coeficientes e, dessa forma, penalizar mais severamente coeficientes com valores pequenos. O peso na regularização proporciona ao método a propriedade de óraculo, portanto, para uma escolha adequada de λ, o *Adaptive Lasso* realiza seleção de variáveis tão bem quanto o oráculo [73].

2.5 Algoritmos de Discretização

É sabido [67] que com o uso de Modelos Lineares Generalizados – que são aditivos, ou seja, as contribuições de cada covariável são sempre adicionadas à função estimada – as covariáveis contínuas podem, para valores positivos, sempre aumentar o valor da variável resposta (crescente) ou diminuir (descrescente). Em outras palavras, o comportamento da variável resposta em relação às covariáveis é monótono.

No caso de contribuições não-monótonas – onde, por exemplo, para dada covariável a variável resposta cresce até determinado valor e, em seguida, decresce – o uso de técnicas de discretização se mostra essencial, pois torna possível separar dada covariável original não-monótona em várias covariáveis derivadas monótonas que representam a covariável original em faixas de valores diferentes [69]. É útil observar que, no exemplo citado, tem-se uma discretização supervisionada, pois observa-se o valor da variável resposta para serem definidos as faixas de discretização.

Em pesquisas recentes [28], pode-se citar três algoritmos com ótimo desempenho em geral, a saber: CAIM (*Class-Attribute Interdependence Maximization*) [45], MDLP (*Minimum Description Length Principle*) [37] e ModChi2 (*Modified Chi2*)[64].

2.5.1 CAIM

O CAIM (*Class-Attribute Interdependence Maximization*) [45] é um algoritmo de discretização que busca maximizar uma métrica definida como interdependência classe-atributo, que mede a interdependência entre a variável resposta e as covariáveis discre-

tizadas. O algoritmo utiliza método guloso, buscando máximos locais, executando-se o algoritmo até que em dada iteração a métrica calculada seja menor que a calculada na iteração anterior – além disso, o método considera que cada covariável precisa ter no mínimo um intervalo para cada classe diferente.

O CAIM é um método de discretização com as seguintes características:

- estático: gera a discretização antes de algum método de aprendizagem de máquina;
- global: utiliza todas as observações para gerar as faixas de discretização;
- incremental: inicia com uma discretização simples e passa por um processo de refinamento, não sendo necessário determinar a quantidade de faixas;
- supervisionado: leva em consideração os valores da variável reposta ao discretizar as covariáveis;
- *top-down*: a lista de pontos de corte para a discretização é inicialmente vazia e ganha novos pontos a cada passo.

É útil citar ainda que estudos recentes [28] indicam o CAIM como um dos algoritmos de discretização mais simples e efetivos.

2.5.2 MDLP

O MDLP (*Minimum Description Length Principle*) [37] é um método de discretização que utiliza a heurística de minimização de Entropia para discretizar covariáveis contínuas em faixas. O algoritmo tem como critério de parada uma métrica baseada no *Minimum Description Length Principle*, calculada para observar o ganho de uma nova partição.

Assim como o CACC, o MDLP é um algoritmo de discretização estático, incremental, supervisionado e *top-down*, no entanto, não é global, mas sim local, pois utiliza apenas observações locais para decidir as faixas de discretização.

É válido apontar que pesquisas recentes [28] demonstram que o MDLP em geral fornece um *tradeoff* satisfatório entre o número de faixas de discretização e a acurácia.

2.5.3 ModChi2

O ModChi2 (*Modified Chi2*) [64] é um método de discretização baseado no teste χ^2 [21] para definir quando valores colocados em faixas serão agrupados. O algoritmo tem como critério de parada o nível de consistência após cada passo da discretização. Tal algoritmo é uma evolução de seus antecessores *ChiMerge* [41] e *Chi*2 [47], que busca maior acurácia e consitência dos dados após a discretização.

Assim como o CACC, o ModChi2 é um algoritmo de discretização estático, global, incremental e supervisionado, no entanto, não é *top-down*, mas sim *bottom-Up*, pois a lista de pontos de corte para discretização é inicialmente igual a todos os pontos e tem seus valores agrupados a cada etapa.

2.6 Avaliação de Modelos de Regressão

Com o uso de métodos de regressão, dois pontos principais podem ser observados numa etapa posterior de avaliação. Primeiro, é necessário avaliar se os modelos ajustados produzem resultados confiáveis e, segundo, dado um conjunto de modelos válidos, deve-se compará-los de forma a definir qual o modelo com melhores resultados. Dessa forma, serão apresentadas algumas técnicas de verificação de qualidade de ajuste e de comparação de modelos.

2.6.1 Teste de Wald

Em métodos com resposta binomial, para verificar se os coeficientes β_i encontrados no modelo ajustado para cada covariável X_i são estatisticamente significativos, é realizado o teste de Wald [32]. A hipótese nula do teste de hipótese é definida como:

$$H_0 : \beta_i = 0 \tag{2.11}$$

O teste de Wald é então obtido através da comparação entre a estimativa de máxima verossimilhança do parâmetro $\hat{\beta}_i$ e a estimativa de seu erro padrão. A razão resultante, sob a hipótese nula acima tem distribuição normal padrão Z, ou seja:

$$W = \frac{\hat{\beta}_i}{\widehat{se}\left(\hat{\beta}_i\right)} \sim \mathcal{N}(0,1) \tag{2.12}$$

onde $\widehat{se}$ representa a estimativa do erro padrão.

Assim, para um dado nível de significância α, realizando o teste de Wald, o p-valor menor que α rejeita a hipótese nula mencionada, ou seja, o coeficiente β_i encontrado no modelo ajustado para dada covariável X_i é diferente de zero, logo, a covariável é significativa no modelo.

2.6.2 Teste de Hosmer-Lemeshow

Com a finalidade de verificar a qualiddade do ajuste de um modelo, pode-se utilizar o teste de Hosmer-Lemeshow [35]. Este teste avalia o modelo gerado a partir da comparação

entre os resultados estimados e observados. A amostra utilizada na avaliação é dividida em g grupos de tamanho aproximadamente iguais segundo suas respostas estimadas ordenadas – ou seja, caso g seja igual a 4, tem-se um grupo para as observações com resposta até 0.25, um segundo grupo com resposta entre 0.25 e 0.50, um terceiro entre 0.50 e 0.75, e um quarto com resposta entre 0.75 e 1.00. Hosmer e Lemeshow propõem o uso de $g = 10$.

Com os resultados para cada grupo separado, a estatística de Hosmer e Lemeshow é obtida comparando as respostas estimadas e observadas para cada grupo. Tem-se, a partir da teoria de máxima verossimilhança, que o valor observado na estatística deve seguir aproximadamente uma distribuição qui-quadrado com $g - 2$ graus de liberdade para que o modelo em análise seja considerado ajustado [35]. Logo, realiza-se o teste com hipótese nula de que o modelo está bem ajustado e, caso o p-valor seja maior que um nível de significância α definido, a hipótese nula não é rejeitada, corroborando a validade do modelo.

2.6.3 Colinearidade

Colinearidade ou multicolinearidade ocorre quando duas ou mais covariáveis são aproximadamente determinadas pela combinação linear de outras covariáveis do modelo [38]. Com colinearidade perfeita, onde uma covariável é combinação linear perfeita de outras, é impossível obter uma estimativa única dos coeficientes de regressão com todas as covariáveis no modelo [36]. Nesse caso, os desvios padrão dos coeficientes tendem a ser altos ou inflados, o que pode ocasionar estimativas de coeficiente não confiáveis.

Duas medidas usuais para diagnosticar colinearidade são tolerância e fator de inflação da variância (VIF) [38]. A tolerância indica quanta colinearidade a análise de regressão suporta, enquanto VIF mede quanto da inflação do desvio padrão pode estar sendo causada por colinearidade. A tolerância (T_j) de uma dada covariável é dada por:

$$T_j = 1 - R_j^2 \tag{2.13}$$

onde R_j^2 é o coeficiente de determinação – igual ao quadrado do coeficiente de correlação de Pearson – calculado para a regressão de X_j como resposta sobre as outras covariáveis.

Já o fator de inflação da variância VIF_j correspondente é dado por:

$$VIF_j = \frac{1}{T_j} = \frac{1}{1 - R_j^2} \tag{2.14}$$

Caso todas as covariáveis sejam completamente não correlacionadas umas com as outras, ambos tolerância e VIF possuirão valor 1 – fato facilmente observável nas equação apresentadas, pois R_j^2 tende a 0 para covariáveis não correlacionadas. Pelo mesmo motivo, analisando as equações, vê-se que caso uma covariável esteja muito relacionada a outra(s)

covariável(is), a tolerância tenderá a 0 e a VIF assumirá valores elevados. Assim, por definição, valores muito baixos de tolerância e valores muito altos de VIF quase certamente indicam problemas de multicolinearidade.

2.6.4 Métricas de Validação

Apesar da regressão objetivar a geração de uma resposta contínua, a definição de um ponto de corte para separação da resposta em grupos ou classes pode ser útil para a validação do desempenho de um modelo de regressão. Usualmente, um ponto de corte utilizado é 0.5, onde, por exemplo, indivíduos com resposta maior que 0.5 sejam classificados como "Corruptos" e aqueles com resposta menor ou igual a 0.5 sejam classificados como "NÃO Corruptos".

Após o ajuste de um modelo e a determinação de um ponto de corte, algumas métricas tomam por base nas suas definições os valores descritos em uma matriz de confusão, como mostrado a seguir na Tabela 2.1.

Tabela 2.1: Matriz de Confusão

	Previsão Positiva	**Previsão Negativa**
Observação Positiva	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Observação Negativa	Falto Positivo (FP)	Verdadeiro Negativo (VN)

No caso exemplificado de classificação das respostas como "Corrupto" e "NÃO Corrupto", o resultado positivo pode ser visto como aquele pertencente à classe "Corrupto", enquanto que o negativo é o equivalente à classe "NÃO Corrupto".

Dessa forma, a partir da matriz de confusão, no exemplo citado, tem-se que VP corresponde a observações originalmente da classe "Corrupto" e previstas pelo modelo também como pertencentes à classe "Corrupto". Da mesma forma, observações VN são aquelas realmente da classe "NÃO Corrupto" que também foram classificadas como tal. Já FP são registros inicialmente da classe "NÃO Corrupto" que foi previsto como "Corrupto" e FN são observações originalmente da classe "Corrupto" indicadas pelo modelo como da classe "NÃO Corrupto".

Diversas métricas podem ser extraídas dos valores VP, VN, FP e FN da matriz de confusão, no entanto, apenas as seguintes foram consideradas como adequadas ao presente trabalho: sensibilidade, especificidade, precisão, acurácia e *F-measure*. Em [61], tais métricas são explicadas com mais detalhes.

Sensibilidade

A métrica denominada sensibilidade é definida utilizando os valores VP e FN da matriz de confusão, da forma mostrada pela equação que segue:

$$Sensibilidade = \frac{VP}{VP + FN} \tag{2.15}$$

Assim, a sensibilidade mede do total de originalmente positivos observados, a porcentagem de corretamente previstos como positivos.

Especificidade

A especificidade é definida utilizando os valores VN e FP da matriz de confusão, como mostrado pela equação que segue:

$$Especificidade = \frac{VN}{FP + VN} \tag{2.16}$$

Dessa forma, a especificidade mede do total de originalmente negativos observados, a porcentagem de corretamente previstos como negativos.

Precisão

Define-se a métrica precisão com os valores VP e FP da matriz de confusão, do modo apresentado pela equação que segue:

$$Precisão = \frac{VP}{VP + FP} \tag{2.17}$$

Portanto, a precisão mede do total de positivos previstos por dado modelo, a porcentagem de originalmente positivos, ou seja, qual a acurácia do modelo em termos de classe positiva.

Acurácia

A acurácia pode ser definida a partir dos valores VP, VN, FP e FN da matriz de confusão, como descrito pela equação que segue:

$$Acurácia = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.18}$$

Logo, a acurácia mede do total de observações originalmente positivos e negativos, a porcentagem de previstos corretamente.

F-measure

A métrica *F-measure* pode ser interpretada como uma medida que corresponde à média harmônica da precisão e sensibilidade, da maneira definida na equação a seguir:

$$F\text{-}measure = \frac{(1+\beta^2) \cdot Precisão \cdot Sensibilidade}{\beta^2 \cdot Precisão + Sensibilidade} \tag{2.19}$$

onde β é uma constante real positiva.

Ajustando o valor de β pode-se aplicar maior peso à sensibilidade com $\beta > 1$ ou maior peso à precisão com $0 < \beta < 1$. Usualmente utiliza-se $\beta = 1$ para conferir o mesmo peso para precisão e sensibilidade, métrica comumente chamada de *F1-measure*, resultando na equação seguinte:

$$F1\text{-}measure = \frac{2 \cdot Precisão \cdot Sensibilidade}{Precisão + Sensibilidade} \tag{2.20}$$

Em termos de valores VP, FN e FP da matriz de confusão, a métrica *F-measure* também pode ser escrita através da equação a seguir:

$$F\text{-}measure = \frac{(1+\beta^2) \cdot VP}{(1+\beta^2) \cdot VP + \beta^2 \cdot FN + FP} \tag{2.21}$$

Dessa forma, *F-measure* pode ser utilizada como métrica mais ampla, por considerar tanto precisão quanto sensibilidade. Em situações de classes raras – como, por exemplo, servidores corruptos – pode-se obter um modelo com alta precisão, onde, no entanto, a sensibilidade é baixa, levando a um resultado de pouca abrangência mesmo preciso. Tal cenário pode ser analisado com o uso da métrica *F-measure*, pois ao balancear precisão e sensibilidade, a avaliação é mais completa.

2.7 Trabalhos Correlatos

No que tange aos trabalhos correlatos em relação à presente pesquisa, optou-se por destrinchar um dos principais pontos a serem abordados na solução proposta, a saber: seleção de atributos. Além disso, aplicações de mineração de dados relacionadas com corrupção são indicadas e analisadas sucintamente, com o objetivo de avaliar o que já foi feito na área em estudo.

2.7.1 Seleção de Atributos

Atualmente diversos campos de pesquisa trazem dados com elevado número de atributos a serem analisados na mineração de dados, o que torna a seleção de atributos relevantes um aspecto crucial para resultados satisfatórios em vários trabalhos [63]. Especialmente em trabalhos de bioinformática e medicina, tais cenários de alta dimensionalidade são

frequentemente encontrados [48], principalmente quando são procurados apenas alguns atributos relevantes em meio a conjuntos de dados altamente esparsos – uma pesquisa típica de bioinformática analisa 10^{3-5} atributos em 10^{1-3} observações [48].

Com esse intuito, métodos de regularização são usualmente utilizados para seleção de atributos nos cenários acima mencionados [51] [60] [63] [48]. Normalmente utiliza-se regularização adicionando penalidades aos coeficientes dos atributos analisados na função a ser minimizada, por exemplo, em uma regressão. Três técnicas bastante usadas para regularização [67] [51] são Regressão *Ridge* [34], *Lasso* [66] e *Adaptive Lasso* [73].

Além de regularização, frequentemente são usados outros métodos que selecionam subconjuntos de atributos, ou *Subset Selection*, como, por exemplo, *Stepwise Selection* [67]. No entanto, vê-se que tal técnica muitas vezes apresenta alta variância [67], e assim não reduz o erro de predição do modelo completo, enquanto métodos de regularização são mais contínuos, e não sofrem tanto de alta variabilidade [67].

2.7.2 Corrupção e Mineração de Dados

Na última década, observando tópicos de pesquisa atuais, uma aplicação em voga próxima da avaliação de risco de corrupção é a detecção de fraudes. O principal objetivo da detecção de fraudes é revelar tendências de atos suspeitos. Por exemplo, um tema emergente é usar mineração de dados na detecção de fraudes financeiras. Uma revisão da literatura acadêmica de tal aplicação [50] mostra a sua utilização bem sucedida em detecção de fraudes de cartão de crédito, lavagem de dinheiro, previsão de falência, entre outros. Esta revisão também identifica técnicas de mineração de dados comuns usadas na detecção de fraudes, incluindo Redes Neurais Artificiais [33], Árvores de Decisão [69] e Regressão Logística [36]. Nesse contexto, há ainda uma recente pesquisa no tema de detecção de fraudes baseada em mineração de dados [53] que exibe um resumo dos artigos técnicos publicados e revisão sobre o tópico. Esse levantamento, assim como em outros trabalhos [54] [43], inclui comentários sobre aplicações semelhantes, no entanto, não menciona combate à corrupção focado em indivíduos.

Em relação aos aspectos de corrupção, pesquisa relacionada com processos de licitação e contratação públicas também já foi realizada, embora não tão amplamente como na detecção de fraudes. O uso de *clustering* e regras de associação para o problema de cartéis em licitações públicas [22] encontrou resultados que corroboram a aplicação de mineração de dados na prevenção da corrupção. Outro artigo [1] mostra o uso de *Naïve Bayes* para avaliar o risco de corrupção de contratos públicos. Os autores aplicam logaritmo natural da contagem dos parâmetros para obter atributos discretizados e baseiam sua avaliação nos resultados das probabilidades condicionais. Adicionalmente, um artigo recente [58] apresenta o uso de ontologias probabilísticas para projetar e testar um modelo que realiza

a fusão de informações para detectar possíveis fraudes em licitações envolvendo dinheiro federal no Brasil. Outro trabalho [57] usa MEBN (*Multi-Entity Bayesian Networks*) [46] e linguagens de ontologias probabilísticas [56] para criar um modelo que gera conhecimento na área de fraudes de licitações.

Portanto, é possível afirmar que mineração de dados tem sido amplamente utilizada para a detecção de fraudes, e trabalhos para a prevenção da corrupção em processos de licitação e contratos também já foram feitos. No entanto, de nosso conhecimento, métodos para geração de modelos preditivos nunca foram usados para estimar risco de corrupção de indivíduos. Assim, este trabalho de modelagem preditiva com a utilização de bases de dados disponíveis para o Governo Federal configura o primeiro trabalho que usa mineração de dados para medir o risco de corrupção de servidores públicos federais.

Capítulo 3

Solução Proposta

Com o intuito de resolver o problema definido no âmbito deste trabalho, a solução proposta envolve seguir a sequência de fases definidas no CRISP-DM, como explanado na Seção 2.1.1. Dessa forma, o trabalho perpassa várias fases simultaneamente, além de ser necessário muitas vezes retornar a fases anteriores.

Buscando esclarecer a solução proposta referente a sua adequação às fases do CRISP-DM, assim como deixar claro cada aspecto do trabalho, inicialmente tem-se a Figura 3.1 com uma breve indicação das etapas iniciais da solução proposta, a serem detalhadas nas próximas seções.

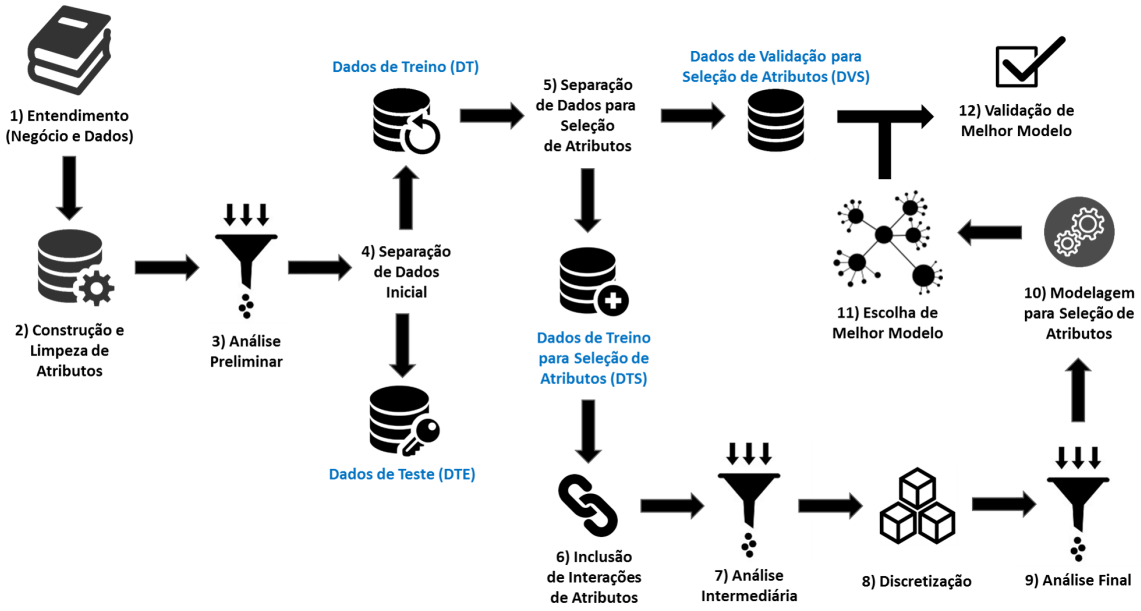

Figura 3.1: Etapas iniciais da solução proposta

Nessa figura os passos numerados são descritos como:

1. **Entendimento (Negócio e Dados)**: refere-se às atividades descritas nas fases de Entendimento do Negócio na Seção 3.1 e Entendimento dos Dados na Seção 3.2;
2. **Construção e Limpeza de Atributos**: obtenção de dados, assim como a limpeza dos mesmos e a construção dos atributos a serem utilizados, apresentadas nas Seções 3.3.1 e 3.3.2;
3. **Análise Preliminar**: etapa de análise de variância e correlação dos atributos construídos, a ser apresentada na Seção 3.3.3;
4. **Separação de Dados Inicial**: dados são separados em treino (DT) e teste (DTE), como será descrito na Seção 3.3.4 – vale frisar que tais siglas serão utilizadas nas seções seguintes para melhor entendimento;
5. **Separação de Dados para Seleção de Atributos**: para uso na seleção de atributos DT é separado em treino (DTS) e validação (DVS), detalhado na Seção 3.3.5;
6. **Inclusão de Interações de Atributos**: adição de interações dos atributos construídos, conforme será explanado na Seção 3.3.5.1;
7. **Análise Intermediária**: nova análise de variância e correlação, adicionalmente a de não linearidade, descrita na Seção 3.3.5.2;
8. **Discretização**: dados são discretizados com diferentes algoritmos, como será delineado na Seção 3.3.5.3;
9. **Análise Final**: os dados discretizados por cada um dos algoritmos passam por análise final de variância e correlação, conforme descrito na Seção 3.3.5.3;
10. **Modelagem para Seleção de Atributos**: é realizada criação de modelos com uso de regularização para cada um dos conjuntos de dados com discretizações com algoritmos diferentes, de forma a selecionar atributos relevantes em cada modelagem, como será visto na Seção 3.4.1;
11. **Escolha do Melhor Modelo**: os modelos criados para cada conjunto de dados são avaliados a partir de seus resultados nos próprios dados utilizados para modelagem (DTS) e é selecionado o melhor modelo, conforme será apresentado na Seção 3.4.1;
12. **Validação de Melhor Modelo**: com o melhor modelo escolhido, o mesmo será aplicado no conjunto de dados DVS para atestar a validade do modelo selecionado, como será delineado na Seção 3.4.1.

É necessário observar que o melhor modelo já validado é de utilidade apenas para selecionar atributos, pois o mesmo não será o modelo final. Pode-se perceber também

que há dados separados inicialmente ainda não utilizados, a saber: dados de teste (DTE). Tanto as informações do melhor modelo quanto os conjuntos DT e DTE serão úteis nas fases seguintes.

Para delinear os passos finais, a Figura 3.2 apresenta uma visão geral das atividades realizadas.

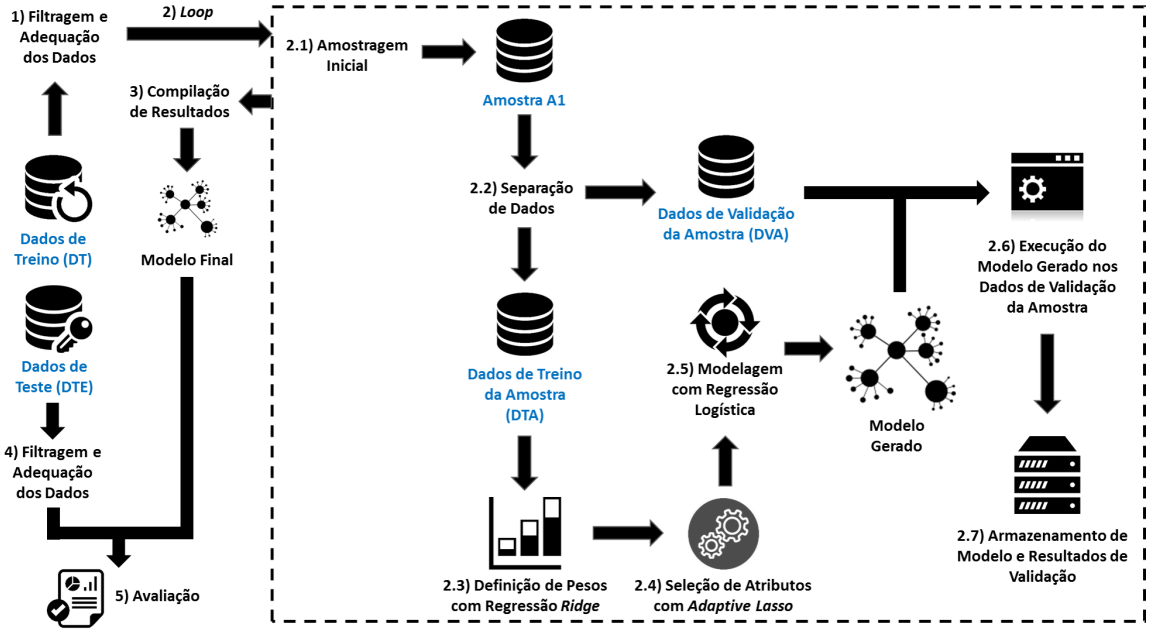

Figura 3.2: Etapas finais da solução proposta

Em tal figura, tem-se os seguintes passos:

1. **Filtragem e Adequação dos Dados**: o conjunto DT criado nas etapas iniciais tem seus atributos filtrados e ajustados de modo a deixá-lo com o mesmo formato dos atributos selecionados pelo melhor modelo obtido nas etapas iniciais, como apresentado na Seção 3.4.1;

2. ***Loop***: são executados dois *loops*, o primeiro com 5.000 iterações e o segundo com 1.000 iterações, sendo que a única diferença é que no segundo são ignorados o terceiro e quarto passo. A estrutura geral do *loop*, detalhado na Seção 3.4.2, é:

   2.1. **Amostragem Inicial**: é obtida uma amostra aleatória A1 do conjunto DT criado nas etapas iniciais;
   2.2. **Separação de Dados**: a amostra A1 é separada em treino (DTA) e Validação (DVA);
   2.3. **Definição de Pesos com Regressão *Ridge***: é construído um vetor de pesos a partir dos coeficientes da Regressão *Ridge* aplicada em DTA;
   2.4. **Seleção de Atributos com *Adaptive Lasso***: com o vetor de pesos definido

no passo anterior, executa-se *Adaptive Lasso* em DTA;

2.5. **Modelagem com Regressão Logística**: a partir dos atributos escolhidos com o *Adaptive Lasso*, é gerado um modelo de regressão logística com DTA;

2.6. **Execução do Modelo Gerado nos Dados de Validação da Amostra**: o modelo gerado no passo anterior é aplicado no conjunto de dados DVA;

2.7. **Armazenamento de Modelo e Resultados de Validação**: os dados resultantes da aplicação do modelo gerado em DVA são armazenados.

3. **Compilação de Resultados**: após as duas execuções do *loop*, os resultados são compilados de modo a criar um modelo final como combinação dos modelos gerados no segundo *loop*;

4. **Filtragem e Adequação dos Dados**: o conjunto DTE tem seus atributos filtrados e ajustados de modo a deixá-lo com o mesmo formato dos atributos selecionados no modelo final obtido no passo anterior;

5. **Avaliação**: Como será apresentado na Seção 4.1, diversas atividades de avaliação são realizadas no modelo final para atestar sua validade, como, por exemplo, sua execução em DTE;

Vale atentar ainda que a última fase do CRISP-DM, Implantação, será delineada na Seção 4.2.

3.1 Entendimento do Negócio

Objetivando conhecer o domínio do presente trabalho, compreendendo o funcionamento da Controladoria-Geral da União (CGU) no que tange a combate à corrupção e aos problemas relacionados a sua área de atuação, principalmente em relação a corruptibilidade de servidores públicos federais, o Entendimento de Negócio foi dividido em duas subseções, a saber: Combate à Corrupção, com uma visão geral do trabalho realizado pela CGU e mais especificamente pela DIE; e Contexto do Trabalho, delineando facilitadores, restrições e objetivos do presente trabalho.

3.1.1 Combate à Corrupção

A CGU atualmente exerce as atividades de órgão central do sistema de controle interno do Poder Executivo federal. Por meio de sua Secretaria Federal de Controle Interno (SFC), a CGU fiscaliza e avalia a execução de programas de governo, inclusive ações descentralizadas a entes públicos e privados realizadas com recursos oriundos dos orçamentos da União; realiza auditorias e avalia os resultados da gestão dos administradores públicos

federais; apura denúncias e representações; exerce o controle das operações de crédito; e, também, executa atividades de apoio ao controle externo.

Dentre outras atribuições [10], compete à CGU assistir direta e imediatamente ao Presidente da República no desempenho de suas atribuições quanto aos assuntos e providências que, no âmbito do Poder Executivo federal, sejam atinentes à prevenção e ao combate à corrupção. A CGU atua ostensivamente no combate à corrupção e ao desvio de recursos públicos, por meio dos trabalhos de auditoria e fiscalização que realiza.

Já sua Diretoria de Pesquisas e Informações Estratégicas (DIE), órgão ligado à Secretaria-Executiva, possui diversas competências relacionadas ao combate à corrupção, sendo várias delas especificamente com foco em servidores públicos federais. Dentre outras atividades, à DIE compete [16]:

> Coordenar, no âmbito da Controladoria-Geral da União, o atendimento a demandas provenientes da Casa Civil da Presidência da República, visando subsidiar a análise dos nomes indicados para ocupar cargos em comissão no Poder Executivo federal;
>
> Acompanhar e analisar a evolução patrimonial dos agentes públicos do Poder Executivo federal, na forma do Decreto nº 5.483, de 30 de junho de 2005;
>
> Construir cenários para subsidiar de forma estratégica as atividades desenvolvidas pela Controladoria-Geral da União, e antecipar, em situações críticas, o encaminhamento preventivo de soluções e o apoio à tomada de decisão;
>
> Prestar assessoramento ao Secretário-Executivo da Controladoria-Geral da União por meio de coleta, busca e tratamento de informações de natureza estratégica para a atuação da Controladoria-Geral da União, com emprego intensivo de recursos de tecnologia da informação e de atividades de investigação e inteligência.

Dessa forma, a DIE, por meio de ações de produção de informações estratégicas, rotineiramente realiza atividades de investigação de possíveis irregularidades cometidas por servidores públicos federais, de forma a subsidiar a atuação da CGU no combate à corrupção. No entanto, atualmente há aproximadamente 1 milhão e 300 mil servidores públicos federais ativos, aposentados e pensionistas em 214 órgãos da Administração Pública Federal direta, instituições federais de ensino, ex-territórios, autarquias, fundações e empresas públicas[1], todos eles sujeitos à investigação. Assim, devido ao grande número de servidores federais existentes atualmente, a DIE acaba muitas vezes restringindo sua atuação aos casos deflagrados externamente, sem realizar exames sistemáticos na extensão pretendida.

Além disso, no que tange aos esforços de auditoria e fiscalização da SFC, a DIE apoia sua atuação com a geração de informações estratégicas em diferentes níveis de detalhamento. No entanto, como a CGU é um órgão com pólos regionais em todos os estados do Brasil e cujas operações de fiscalização e controle abrangem unidades e

[1] Site de apresentação do Sistema Integrado de Administração de Recursos Humanos (SIAPE) do Governo Federal: http://www.siapenet.gov.br/oque.htm

servidores dos mais diversos órgãos no país, o suporte de informações estratégicas da DIE é na maioria das vezes limitado a casos ou situações específicas, não comportando uma atuação totalmente efetiva e em larga escala.

3.1.2 Contexto do Trabalho

É válido frisar que a DIE possui competências [16] ligadas ao acesso a informações diversas e também relacionadas com monitoramento e soluções com o uso de tecnologia:

> Requisitar dados e informações a agentes, órgãos e entidades públicas e privadas que gerenciem recursos públicos federais para subsidiar a produção de informações estratégicas necessárias ao desenvolvimento das atividades da Controladoria-Geral da União;
>
> Solicitar às unidades da Controladoria-Geral da União dados e informações que subsidiem e complementem atividades de investigação e inteligência;
>
> Prospectar, avaliar e propor soluções de tecnologia para as atividades de pesquisa e investigação na área de produção de informação estratégica;
>
> Realizar monitoramento contínuo dos gastos públicos por meio de técnicas e ferramentas de análise aplicadas às bases de dados governamentais.

Assim, considerando a vasta gama de dados de servidores públicos federais a que a CGU tem acesso e suas competências relacionadas a tecnologia, são objetivos deste projeto a construção de modelos preditivos para avaliação de risco de corrupção de servidores públicos federais e a descoberta de conhecimento no que tange a informações sobre corruptibilidade de servidores públicos federais. Dessa forma, tem-se já em decreto as competências necessárias para tal, além das definições das atividades onde as informações produzidas serão consumidas. Assim, tanto consegue-se apoiar as atividades de auditoria e fiscalização da SFC, quanto produzir informações estratégicas no próprio âmbito de atividades internas à DIE, além de se adquirir maior conhecimento sobre regras e cenários relacionados a combate à corrupção – descoberta de conhecimento que pode subsidiar praticamente todas as atividades da DIE e consequentemente da CGU.

3.2 Entendimento dos Dados

Buscando analisar corruptibilidade de servidores públicos federais, diversas bases de dados a que a DIE tem acesso foram identificadas como úteis para o presente trabalho. Para melhor entendimento dos dados, as informações disponíveis foram divididas em quatro dimensões, a saber:

- Dimensão de Corrupção;
- Dimensão Funcional;

- Dimensão Política; e
- Dimensão de Vínculos Societários.

A seguir, cada dimensão será apresentada, identificando-se os dados existentes e suas possíveis relações com corruptibilidade, e descrevendo as informações encontradas em cada base de dados que foram consideradas relevantes aos objetivos deste trabalho pelos especialistas da DIE. Além da compreensão das informações contidas nos dados, serão identificados também possíveis problemas de qualidade dos dados e restrições de formatação a serem tratadas nas fases seguintes.

3.2.1 Dimensão de Corrupção

Com relação a servidores públicos federais, a própria CGU mantém o Cadastro de Expulsões da Administração Federal (CEAF)[2]. O CEAF é um banco de informações que reúne as penalidades expulsivas aplicadas (demissão, cassação de aposentadoria e destituição de cargo em comissão ou função comissionada), no âmbito do Poder Executivo federal, a servidores civis, efetivos ou não, desde o ano de 2003.

Nesse sentido, demissão é a pena aplicável ao servidor público efetivo que comete infração grave no exercício de cargo e que ainda se encontra nos quadros da Administração Pública Federal. Já a cassação de aposentadoria é a punição aplicada quando o servidor já esta aposentado, mas for penalizado com a demissão por ato praticado enquanto encontrava-se em exercício na Administração Pública. Enquanto que a destituição de cargo em comissão ou função comissionada é a penalidade expulsiva aplicada a pessoa que ocupava somente cargo em comissão ou função comissionada, não sendo servidor público efetivo da Administração Pública Federal.

É importante notar ainda que as penas estão muitas vezes relacionadas com a denominação de conflito de interesses. A Lei nº 12.813 [17], que dispõe sobre o conflito de interesses no exercício de cargo ou emprego do Poder Executivo federal, define em seu art. 3º que conflito de interesses é "a situação gerada pelo confronto entre interesses públicos e privados, que possa comprometer o interesse coletivo ou influenciar, de maneira imprópria, o desempenho da função pública". Ademais a Lei acrescenta no parágrafo 2º do mesmo artigo que "a ocorrência de conflito de interesses independe da existência de lesão ao patrimônio público, bem como do recebimento de qualquer vantagem ou ganho pelo agente público ou por terceiro".

Publicado na página da CGU há ainda um relatório estatístico de punições expulsivas[3], que contempla as informações referentes à totalidade de penalidades aplicadas. Tal

[2]CEAF: Cadastro de Expulsões da Administração Federal – Link: http://www.portaldatransparencia.gov.br/expulsoes/entrada

relatório inclui o gráfico exibido na Figura 3.3, onde é possível observar que a grande maioria das expulsões são fundamentadas por atos relacionados à corrupção.

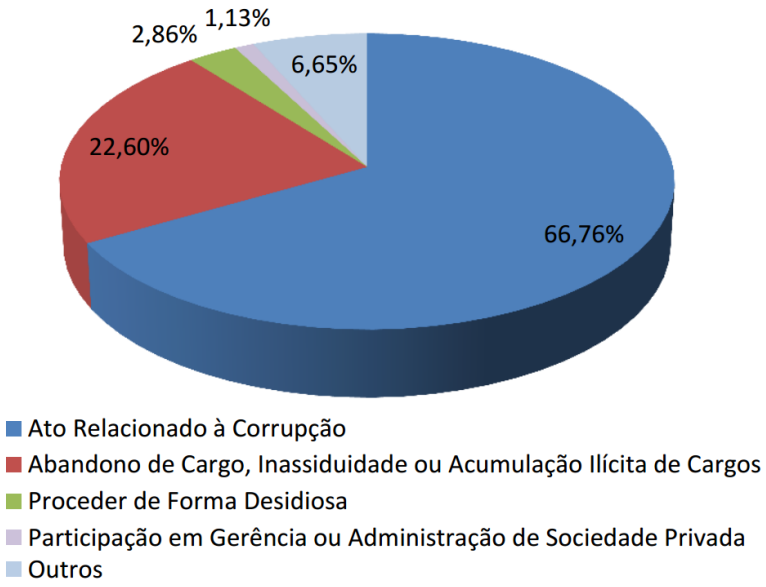

Figura 3.3: Fundamentações das Punições Expulsivas do CEAF[3]

Na base de dados do CEAF é possível filtrar as expulsões pelas fundamentações legais. Utilizando novamente como referência a definição mencionada no Capítulo 1 – amparada pela Lei 8.429 [5] –, tem-se corrupção como sendo o ato de improbidade ou valimento indevido de cargo público que importe enriquecimento ilícito, cause lesão ao erário ou atente contra os princípios da Administração Pública. Dessa forma, filtrando os registros do CEAF pelas fundamentações legais da Lei citada pode-se obter apenas servidores públicos federais considerados corruptos pela definição mencionada.

Portanto, numa escala contínua probabilística de 0 a 1, nos métodos de aprendizagem de máquina supervisionada poderão ser utilizados como servidores públicos federais com risco de corrupção = 1 todos aqueles expulsos cadastrados no CEAF cuja fundamentação legal seja condizente com a Lei nº 8.429 e, por conseguinte, com a definição de corrupção considerada neste trabalho.

Além do campo de fundamentação legal das expulsões e o CPF de cada servidor expulso, para fins de identificação única, tem-se ainda como campo útil a data de publicação da expulsão dos servidores. O campo de fundamentação legal apresenta padronização ajustada, no entanto, tanto CPF quanto data de publicação de expulsão possuem problemas de valores inconsistentes, sendo necessária limpeza na fase de preparação dos dados.

[3]Relatório estatístico de punições expulsivas (CEAF) – Link: http://www.cgu.gov.br/assuntos/atividade-disciplinar/relatorios-de-punicoes-expulsivas/arquivos/punicoes-mensal.pdf

3.2.2 Dimensão Funcional

A dimensão funcional trata dos dados de servidores públicos federais registrados no âmbito de atuação de cada agente, podendo ser relacionados a informações básicas, como tempo de serviço e função, sanções na investidura do cargo, resultantes de práticas de má gestão do dinheiro público – e investigações de ofício ou não por parte da CGU. Assim, a dimensão funcional pode ter suas informações divididas em três grupos de dados: dados básicos, sanções e investigações.

3.2.2.1 Dados Básicos

Os dados funcionais básicos dos aproximadamente 1 milhão e 300 mil servidores públicos federais ativos, aposentados e pensionistas de 214 órgãos da Administração Pública Federal direta, instituições federais de ensino, ex-territórios, autarquias, fundações e empresas públicas são cadastrados no Sistema Integrado de Administração de Recursos Humanos (SIAPE)[3]. O SIAPE mantém dados cadastrais, pessoais, funcionais e de processamento da folha de pagamento de servidores ativos e inativos, pensionistas e aposentados do Governo Federal, e é gerido pela Secretaria de Recursos Humanos do Ministério do Planejamento, Orçamento e Gestão (MPOG), órgão central do Sistema de Pessoal Civil da Administração Federal (SIPEC).

A estrutura de cargos da Administração Pública Federal divide-se entre os cargos efetivos e cargos em comissão. Os cargos efetivos, em regra, são ocupados por servidores de carreira, mediante aprovação em concurso público. Os cargos em comissão são de livre provimento, ou seja, livre nomeação e exoneração de funcionários, sejam eles de carreira ou de fora do serviço público. Os cargos em comissão dividem-se em quatro categorias principais:

- Cargos de livre provimento das agências reguladoras;
- Cargos de direção das instituições federais de ensino superior;
- Cargos de natureza especial (CNEs); e
- Cargos de direção e assessoramento superior (DAS).

No que tange ao uso dos dados do SIAPE para avaliação de corruptibilidade, vê-se que a corrupção é um conjunto variável de práticas que implica em trocas entre quem detém poder decisório e quem detém poder econômico, visando à obtenção de vantagens – ilícitas, ilegais ou ilegítimas – para os indivíduos ou grupos envolvidos. Assim, partindo-se

[3]Site de apresentação do Sistema Integrado de Administração de Recursos Humanos (SIAPE) do Governo Federal: http://www.siapenet.gov.br/oque.htm

da premissa de que os servidores que exercem funções de confiança e aqueles ocupantes de cargos em comissão são os funcionários que detêm o poder decisório e, portanto, estariam suscetíveis às investidas daqueles que detêm o poder econômico, tem-se que o SIAPE afigura-se como o sistema estruturante capaz de identificar quais seriam os postos-chave com maior propensão à pratica da corrupção.

Por escolha dos especialistas da DIE, este trabalho manteve-se utilizando informações funcionais e de pagamento, não incluindo dados pessoais – como sexo, escolaridade e estado civil. Dessa forma, foram extraídas oito informações diferentes, que serão processadas na fase de preparação dos dados para se tornarem um ou mais atributos. São elas:

1. **Salário bruto:** valor total recebido por dado servidor no exercício de cargo ou função. Pode ser extraído do campo valor bruto nos registros de pagamentos.
2. **Quantidade de cargos:** número de cargos diferentes ocupados pelo servidor público federal. Pode ser calculado a partir do número de diferentes valores do campo código de cargo nos dados de cargos.
3. **Quantidade de órgãos de cargo:** número de órgãos diferentes onde cada servidor ocupou um cargo. Pode ser calculado a partir do número de diferentes valores do campo código de órgão nos dados de cargos.
4. **Quantidade de órgãos de função:** número de órgãos diferentes onde cada servidor ocupou uma função. Pode ser calculado a partir do número de diferentes valores do campo código de órgão nos dados de funções.
5. **Quantidade de atividades:** número de atividades distintas das funções que cada servidor ocupou. Pode ser calculado a partir do número de diferentes valores do campo código de atividade nos dados de funções.
6. **Tempo de cargo:** tempo em dias desde a entrada em dado cargo até sua saída ou o presente momento. Pode ser calculado a partir do diferença entre os valores do campo data de ocupação do cargo e os valores do campo data de exclusão – no caso de data de exclusão vazia, pode ser utilizada a função de obter data atual.
7. **Situação funcional:** valor categórico de situação funcional do servidor, com 20 tipos diferentes nas amostras obtidas, como, por exemplo, ativo permanente, aposentado e instituidor de pensão. Pode ser extraído do campo descrição de situação funcional nos dados cadastrais.
8. **Nível de função:** valor categórico da função exercida por servidor público federal, com 82 tipos diferentes nas amostras obtidas, como, por exemplo, DAS101.3,

DAS102.4 e FCI-0001. Pode ser extraído do campo nível da função nos dados de funções.

Das informações acima, vê-se problemas de qualidade nos campos atinentes à informação de tempo de cargo, a saber: data de ocupação do cargo e data de exclusão. Torna-se necessária, portanto, uma limpeza dos dados de tais campos.

3.2.2.2 Sanções

As sanções de servidores públicos federais aplicadas no âmbito de sua atuação funcional abordadas pelo presente trabalho envolvem resultados de contas julgadas irregulares pelo Tribunal de Contas da União (TCU) e emissão de certificados e constatações pela Controladoria-Geral da União (CGU).

Contas Julgadas Irregulares pelo TCU

No âmbito de gestão pública, a prestação de contas é uma forma de governantes e burocracia estatal promoverem a abertura da gestão à inspeção pública, justificando seus atos à sociedade e, consequentemente, sujeitando-se a sanções em caso de abuso, falta ou ilegalidade [6]. Na ausência de garantias de que gestores atuem como legítimos representantes do povo, buscando, por vezes, interesses próprios e/ou de aliados, a prestação de contas permite que se verifique a probidade dos gestores responsáveis pela Administração Pública. Assim, a Lei 8.443 [6] definiu que ao julgar as contas, o TCU avaliará a gestão e decidirá se essas são regulares, regulares com ressalva, ou irregulares.

Mais especificamente, o art. 16, inciso III, da referida Lei esclarece que as contas serão julgadas irregulares, quando comprovada: a omissão no dever de prestar contas; a prática de ato de gestão ilegal, ilegítimo, antieconômico, ou infração à norma legal ou regulamentar de natureza contábil, financeira, orçamentária, operacional ou patrimonial; o dano ao Erário decorrente de ato de gestão ilegítimo ao antieconômico; ou o desfalque ou desvio de dinheiros, bens ou valores públicos.

Os conceitos anteriores se relacionam na medida em que, pela própria definição legal, o julgamento das contas pela irregularidade realizado pelo TCU é uma punição pela atuação não desejável, ilegítima, e, possivelmente corrupta, por parte do gestor, que tenha ou não implicado em dano ao Erário. Dessa forma, as informações geradas por tais julgamentos são de grande valia na avaliação de risco de corrupção.

Através da orientação dos especialistas da DIE, foram extraídas duas informações diferentes, que serão processadas na fase de preparação dos dados para se tornarem um ou mais atributos. São elas:

1. **Número de contas julgadas irregulares:** quantidade de contas julgadas irregulares a que dado servidor público federal está relacionado como gestor. Pode ser calculado a partir do número de diferentes valores do campo de identificação única de julgamento de contas.

2. **Origem de recursos:** valor categórico da origem dos recursos envolvidos no julgamento de contas, com 7 tipos diferentes nas amostras obtidas, como, por exemplo: acordo ou convênio; prestação de contas anual; e auxílio, contribuição ou subvenção. Pode ser extraído do campo descrição da origem de recursos nos dados de contas julgadas irregulares.

Das informações utilizadas a partir da base de dados de contas julgadas irregulares, vê-se problemas no campo de CPF, com a existência de diversos valores inválidos. Torna-se necessária, portanto, uma limpeza dos dados de tais dados.

Constatações e Certificados da CGU

A CGU atua diretamente no combate à corrupção e ao desvio de recursos públicos, por meio dos trabalhos de auditoria e fiscalização que realiza através de sua Secretaria Federal de Controle (SFC). Para a execução desses trabalhos, a CGU utiliza um sistema de informação denominado Novo Ativa [15], onde são registrados os relatórios produzidos e as situações encontradas. Os documentos eletrônicos gerados no sistema Novo Ativa apresentam diretrizes e procedimentos que devem ser executados pelas unidades integrantes do controle interno. Esses documentos são denominados ordens de serviço (OS) e detalham as chamadas ações de controle, que são as atividades a serem realizadas. Segundo o Manual de Elaboração de Relatórios de Controle Interno [20], um dos tipos de registros nos relatórios produzidos pela CGU é a denominada constatação, que representa a indicação de situações indesejáveis identificadas durante a execução da ação de controle, como a existência de dificuldades, equívocos, situações que contrariam normas, critérios técnicos ou administrativos.

No referido manual as constatações podem ser classificadas como falha formal, falha média e falha grave, de acordo com a ocorrência verificada. A falha grave pode representar: omissão no dever de prestar contas (inclusive sonegação de informações necessárias à atuação do Controle Interno); dano ao erário decorrente de ato de gestão ilegítimo ou antieconômico; desfalque ou desvio de dinheiros, bens ou valores públicos; prática de ato de gestão ilegal, ilegítimo, antieconômico, ou infração a norma legal ou regulamentar que tenha potencialidade de causar prejuízos ao erário ou configure grave desvio relativamente aos princípios a que está submetida a Administração Pública. Já a falha média reflete prática de ato de gestão ilegal, ilegítimo, antieconômico, ou infração a norma legal ou

regulamentar não enquadrado nas hipóteses de falha grave, enquanto que a falha formal são outras constatações, não enquadráveis nas hipóteses acima.

Além disso, a CGU emite como produto da realização das auditorias anuais de contas os chamados certificados, que representam a opinião do Sistema de Controle Interno sobre a exatidão e regularidade da gestão em determinado exercício. Nos termos da Portaria 1.161 da CGU [20], os certificados podem ser de três tipos: de regularidade, regularidade com ressalva e irregularidade. O certificado de regularidade é emitido quando as contas expressarem, de forma clara e objetiva, a exatidão dos demonstrativos contábeis, a legalidade, a legitimidade e a economicidade dos atos de gestão do responsável. Já o certificado de regularidade com ressalva será gerado quando forem evidenciadas quaisquer faltas ou impropriedades não abrangidas pelas hipóteses de certificado de irregularidade. Finalmente, o certificado de irregularidade será emitido quando verificada uma ou mais das seguintes ocorrências: omissão no dever de prestar contas; prática de ato de gestão ilegal, ilegítimo, antieconômico, ou infração a norma legal ou regulamentar de natureza contábil, financeira, orçamentária, operacional ou patrimonial, que tenham potencialidade de causar prejuízos ao erário ou configurem grave desvio relativamente aos princípios a que está submetida a Administração Pública; dano ao erário decorrente de ato de gestão ilegítimo ou antieconômico; desfalque ou desvio de dinheiros, bens ou valores públicos.

Assim, a verificação de constatações com ocorrência de falhas e suas respectivas classificações pode trazer informações intimamente relacionadas à corruptibilidade, já que engloba situações de improbidade de servidores no que tange à má gestão pública ou até desvio de recursos. Da mesma forma, os certificados gerados pela CGU trazem uma compilação da análise de determinada gestão e consequentemente dos servidores públicos responsáveis. Portanto, tais informações podem trazer resultados na avaliação de risco de corrupção.

Com o direcionamento dos especialistas da DIE, seis informações diferentes foram elencadas, a serem futuramente processadas na fase de preparação dos dados para se tornarem um ou mais atributos. São elas:

1. **Número de certificados de regularidade com ressalva:** quantidade de certificados de regularidade com ressalva emitidos que cada servidor público federal encontra-se como responsável pela gestão. Pode ser calculado a partir do número de diferentes valores do identificador único classificação de gestão de responsável, após filtro de tipo de classificação igual a certificado de regularidade com ressalva.

2. **Número de certificados de irregularidade:** quantidade de certificados de irregularidade emitidos que cada servidor público federal encontra-se como responsável pela gestão. Pode ser calculado a partir do número de diferentes valores do

identificador único de classificação de gestão de responsável, após filtro de tipo de classificação igual a certificado de irregularidade.

3. **Quantidade de constatações:** número de constatações onde cada servidor público federal encontra-se como responsável. Pode ser calculado a partir do número de diferentes valores do identificador único de constatação nos dados de responsáveis por constatação.

4. **Quantidade de constatações por OS:** Número de constatações dividido pelo número de OS onde cada servidor público federal encontra-se como responsável. Pode ser calculado a partir do número de diferentes valores dos identificadores únicos de constatação e OS nos dados de responsáveis por constatação e OS.

5. **Quantidade de OS:** número de OS onde cada servidor público federal encontra-se como responsável. Pode ser calculado a partir do número de diferentes valores do identificador único de OS.

6. **Tipo da constatação:** valor categórico da classificação da constatação, com 14 tipos diferentes nas amostras obtidas, como, por exemplo, falha média e falha grave. Pode ser extraído do campo descrição de tipo de constatação nos dados de constatações.

Das informações acima, vê-se problemas de qualidade no campo de tipo da constatação, onde não há padronização de valores, sendo necessária, portanto, uma limpeza dos dados.

3.2.2.3 Investigações

As investigações de servidores públicos federais aplicadas no âmbito de sua atuação funcional disponíveis à DIE compreendem informações de processos administrativos disciplinares e registros de investigações pontuais de ofício ou não realizadas no âmbito da própria DIE.

Processos Disciplinares

As informações sobre Procedimentos Administrativos Disciplinares (PAD), tanto finalizados quanto em curso, instaurados no âmbito dos órgãos, entidades, empresas públicas e sociedades de economia mista do Poder Executivo federal são registradas no Sistema de Gestão de Processos Disciplinares (CGU-PAD)[4] O sistema utiliza informações fornecidas por todo o Poder Executivo federal de modo a concentrar dados de procedimentos disciplinares com informações datadas desde 1992.

[4]Sistema de Gestão de Processos Disciplinares (CGU-PAD) – Link: http://www.cgu.gov.br/assuntos/atividade-disciplinar/cgu-pad

Os dados presentes no CGU-PAD abrangem as diversas fases dos procedimentos administrativos, quais sejam: instauração/instrução; indiciamento/citação/defesa escrita/-relatório final; encaminhado para julgamento; e processo julgado. Após o julgamento dos processos, podem ser aplicadas penalidades, como suspensão, demissão, advertência, destituição de cargo em comissão ou cassação de aposentadoria.

O servidor público federal que comete irregularidades no exercício de suas atribuições pode ter que responder pelos atos nas instâncias civil, penal e administrativa, conforme art. 121 da Lei 8.112 [4]. Assim, a prática de condutas vedadas nos diversos regramentos administrativos públicos ou o descumprimento dos deveres funcionais dão margem à responsabilização administrativa do servidor. Já a causa de danos patrimoniais a terceiros ou à Administração Pública implicam em responsabilização civil, enquanto que a prática de crimes e contravenções ensejam a responsabilidade penal. É válido observar ainda que, no caso de responsabilização civil, a obtenção do ressarcimento poderá ocorrer mediante Tomada de Contas Especial (TCE), atualmente regulamentada pela Instrução Normativa TCU nº 71/2012 [26], destinada à apuração de responsabilidade pelos danos causados à Administração Pública Federal e à obtenção do respectivo ressarcimento.

Portanto, os registros constantes na base de procedimentos administrativos disciplinares centralizada na CGU estão relacionados com informações ensejadoras de cenários de corrupção. Com a vantagem de imputar penalidades diretamente a servidores públicos federais, as informações de julgamento disciplinar e responsabilização podem ser úteis para análise de corruptibilidade.

No entanto, apesar da existência dos dados de responsabilizações civil e criminal, assim como a instauração de TCE e ocorrência de improbidade, no sistema CGU-PAD, devido ao fato dessas informações estarem intimamente ligadas às expulsões no âmbito federal cadastradas no CEAF, o uso de tais informações pode levar a criação de atributos aparentemente relevantes, mas que na verdade não ajudam a avaliação de risco deste trabalho, pois processos disciplinares com demissão, destituição de cargo e cassação de aposentadoria não são influenciadores de expulsão, mas sim a própria expulsão. Assim, a partir da análise dos especialistas da DIE, por não terem relação causal com a expulsão, foram definidas apenas duas informações diferentes para dados dos procedimentos disciplinares, a serem futuramente processadas na fase de preparação dos dados para se tornarem um ou mais atributos. São elas:

1. **Quantidade de penalidades de advertência:** número de procedimentos administrativos disciplinares julgados onde cada servidor público federal recebeu a penalidade de advertência. Pode ser calculado a partir da contabilização do número de valores do campo identificador único de PAD, filtrando o campo tipo de penalidade com o valor advertência.

2. **Quantidade de penalidades de suspensão:** número de procedimentos administrativos disciplinares julgados onde cada servidor público federal recebeu a penalidade de suspensão. Pode ser calculado a partir da contabilização do número de valores do campo identificador único de PAD, filtrando o campo tipo de penalidade com o valor suspensão.

Das informações acima, vê-se problemas no campo de tipo de penalidade devido a inexistência de padrão. Torna-se necessária, portanto, uma limpeza dos dados de tal campo.

Investigações da DIE

A Diretoria de Pesquisas e Informações Estratégicas (DIE) encontra-se, na estrutura da CGU, vinculada à Secretaria-Executiva. Sua criação se deu em virtude da necessidade de produzir informações de interesse da alta administração da CGU e das áreas finalísticas. A DIE é composta atualmente por duas Coordenações-Gerais [16], que realizam atividades complementares. A Coordenação-Geral do Observatório da Despesa Pública (CGODP) é responsável pela aplicação intensiva de recursos de tecnologia da informação para integrar as informações provenientes de diversas bases de dados disponíveis, construindo ferramentas capazes de auxiliar a pesquisa e a produção de conhecimentos sobre tais informações. Já a Coordenação-Geral de Informações Estratégicas (CGIE) é uma unidade voltada a investigações e pesquisas específicas, de forma a subsidiar principalmente as ações de controle e correição.

A CGIE, em auxilio à sua atividade de produzir informações estratégicas, conta com um sistema interno à diretoria denominado Registros de Investigações (RI), onde são armazenadas e controladas as informações sobre as investigações realizadas no âmbito da Coordenação, incluindo diversos dados sobre investigados e outros envolvidos, assim como relatórios não estruturados descrevendo os resultados. As ações da CGIE são desencadeadas a partir de denúncias, pedidos de operações federais deflagradas, dados disponibilizados pela CGODP, dentre outras origens.

Assim, considerando a atuação passada da CGIE em investigações com origens diversas, a presença de servidores públicos federais na base de dados do sistema interno RI pode trazer indicativos em relação à corruptibilidade, já que engloba situações de servidores que, por estarem ou já tiverem passado por uma investigação, possuem indícios de práticas relacionadas a corrupção e estão envolvidos em situação de risco, como denúncias e operações especiais.

Apoiando-se no conhecimento dos especialistas da CGIE, duas informações diferentes foram elencadas, a serem futuramente processadas na fase de preparação dos dados para se tornarem um ou mais atributos. São elas:

1. **Quantidade de investigações:** número de procedimentos de investigação realizados pela CGIE onde cada servidor público federal consta como envolvido. Pode ser calculado a partir da contabilização do número de valores do campo identificador único de investigação nos dados do RI.

2. **Tipo de envolvimento:** valor categórico da classificação do tipo de envolvimento de dado indivíduo em uma investigação, com dois tipos diferentes nas amostras obtidas, a saber: investigado e outro. Pode ser extraído do campo tipo de agente nos dados do RI.

3.2.3 Dimensão Política

A dimensão política trata dos dados de servidores públicos federais relacionados com sua atuação política, nas mais diversas esferas. Assim, esta dimensão pode ter suas informações divididas em dois grupos de dados: filiação partidária e candidatura eleitoral.

3.2.3.1 Filiação Partidária

A filiação partidária é o ato pelo qual um eleitor aceita e adota um programa, e passa a integrar um partido político. Nos termos do art. 16 da Lei dos Partidos Políticos [7], só pode filiar-se a partido o eleitor que estiver no pleno gozo de seus direitos políticos. Ao integrar um partido político, o filiado segue o disposto em estatuto próprio do respectivo partido. Ainda conforme dispõe a referida Lei, em seu art. 5º, o estatuto define como é exercida a ação do partido, além de estabelecer as normas de disciplina e fidelidade partidária dos filiados, sendo registrado no Tribunal Superior Eleitoral (TSE) após o partido adquirir personalidade jurídica.

Segundo cada estatuto partidário, os filiados têm o direito de participar ativamente das atividades do Partido e manifestar seus pontos de vista nas reuniões. Além disso, aqueles que se filiarem devem votar em candidatos do Partido, apoiar, empenhar-se e participar assiduamente das campanhas políticas e eleitorais dos candidatos do Partido, assim como manter relações de urbanidade e respeito com os detentores de mandatos eletivos. Outro ponto importante versa sobre o exercício de cargos na Administração Pública – aspecto presente em diversos estatutos. Segundo os documentos, é expresso que os filiados investidos em cargos de confiança na Administração Pública, direta ou indireta, deverão exercê-los com probidade, fidelidade aos princípios programáticos e à orientação do Partido, sendo obrigados a prestar contas de suas atividades, caso sejam convocados.

Tendo o exposto em vista, vê-se que o relacionamento obrigatório, definido em estatuto partidário, entre filiados servidores públicos e políticos pode levar ao desempenho de atividades públicas com viés político em detrimento do interesse público. Assim, rela-

ções sociais de caráter pessoal tendem a aumentar a influência política na Administração Pública, dando margem a atividades corruptas. Portanto, a partir da clara abertura de possibilidades de corrupção por influência política arbitrária na Administração Pública com fins adversos ao interesse público, considerou-se válido abordar informações de filiação político-partidária de servidores públicos federais neste trabalho.

Ainda em relação às informações de filiação, faz-se necessário entender como funcionam os processos de desfiliação e cancelamento de filiação. Na desfiliação, para desligar-se de seu partido político, o filiado deve fazer comunicação escrita ao órgão de direção municipal ou zonal e ao juiz da zona eleitoral onde for inscrito. Passados dois dias da entrega da comunicação ao cartório eleitoral, o vínculo se extinguirá para todos os efeitos – conforme dispõe a Lei nº 9.096 [7], art. 21, caput, e parágrafo único. Importante observar que deve haver apenas uma única filiação por cidadão, de forma que detectada eventual duplicidade de filiação – após notificação do filiado e partidos envolvidos pela Justiça eleitoral – caso não haja comprovação da inexistência da filiação ou de regular desfiliação, juiz poderá declarar a nulidade de ambas as filiações ou o sistema de filiações atualizará a situação das filiações automaticamente para canceladas, consoante prevê o parágrafo único do art. 22 da Lei nº 9.096 [7]. Além do cancelamento em casos de duplicidade, por meio judicial ou via sistema, a referida Lei, em seu art. 22, incisos I a IV, prevê também que a filiação partidária poderá ser cancelada a pedido do partido político nos casos de morte, perda dos direitos políticos, expulsão e outras formas previstas no estatuto dos partidos políticos.

Levando em consideração o conhecimento dos especialistas da DIE, a partir dos dados fornecidos pelo TSE foram selecionadas seis informações diferentes no âmbito de filiação, a serem futuramente processadas na fase de preparação dos dados para se tornarem um ou mais atributos. São elas:

1. **Número de cancelamentos:** quantidade de ocorrências de cancelamentos de filiações a partidos registradas para cada servidor público federal. Pode ser calculada a partir da contabilização de quantidade de valores não vazios do campo data de cancelamento nos dados do filiação do TSE.

2. **Número de desfiliações:** quantidade de ocorrências de desfiliação de partidos registradas para cada servidor público federal. Pode ser calculada a partir da contabilização de quantidade de valores não vazios do campo data de desfiliação nos dados do filiação do TSE.

3. **Tempo de filiação:** número de dias de filiação partidária para cada servidor público federal. Pode ser calculado a partir da diferença entre o campo data de filiação e data de desfiliação ou data de cancelamento nos dados do filiação do TSE.

4. **Motivo de cancelamento:** valor categórico da classificação do motivo de cancelamento de filiação de dado servidor, com cinco tipos diferentes nas amostras obtidas, como, por exemplo: cancelamento automático, judicial ou a pedido do partido. Pode ser extraído do campo descrição de motivo de cancelamento nos dados de filiação do TSE.

5. **Sigla do partido:** valor categórico de sigla de partido a qual cada servidor público federal encontra-se filiado, com 30 tipos diferentes nas amostras obtidas, como, por exemplo: PT, PSDB e DEM. Pode ser extraído do campo sigla do partido nos dados de filiação do TSE.

6. **Quantidade de filiações:** número de ocorrências de filiação a partidos registradas para cada servidor público federal. Pode ser calculado a partir da contabilização de quantidade de valores diferentes de sigla do partido em conjunto com data de filiação para cada servidor nos dados de filiação do TSE.

Das informações utilizadas a partir das bases de dados do TSE, vê-se várias necessidades de limpeza de dados, como inconsistências entre os campos – data de desfiliação ou cancelamento anterior à data de filiação – ou erros isolados, como datas inválidas.

3.2.3.2 Candidatura Eleitoral

A elegibilidade para candidatura eleitoral é matéria tratada em nosso sistema jurídico em nível constitucional. De acordo com o art. 14, parágrafo 3º, da Constituição Federal [3] as condições de elegibilidade são: a) nacionalidade brasileira; b) pleno gozo dos direitos políticos; c) alistamento eleitoral; d) domicílio eleitoral na circunscrição da eleição; e) filiação partidária; f) ter a idade mínima exigida. Dessa forma, servidores públicos que satisfizerem tais requisitos também podem ser candidatos para cargos eletivos.

Nesse sentido, para candidatar-se a cargo eletivo, o servidor público federal possui o direito de concessão de "licença para atividade política", sendo afastado a partir do dia imediato ao do registro de sua candidatura perante a Justiça Eleitoral, até o décimo dia seguinte ao do pleito, conforme redação dada pela Lei nº 9.527 [9]. A partir do registro da candidatura e até o décimo dia seguinte ao da eleição, o servidor fará jus à licença, assegurados os vencimentos do cargo efetivo, somente pelo período de três meses.

É válido atentar que os próprios partidos políticos, por ocasião do registro das candidaturas, fixam e informam, em formulário próprio da Justiça Eleitoral, os valores máximos de gastos de campanha, por cargo eletivo. Sendo que gastos além dos limites máximos estabelecidos pelo partido político sujeita os responsáveis ao pagamento de multa no valor de cinco a dez vezes a quantia em excesso, podendo os responsáveis responder por abuso

do poder econômico e, em razão disso, ter cassado o seu registro ou o diploma, se este já houver sido outorgado, além de ficar inelegível pelo prazo de oito anos.

Dessa forma, a candidatura eleitoral soma-se ao cenário de filiação partidária, de forma a representar relacionamentos ainda mais estreitos entre o servidor público federal candidato e a seara política, ao se tornar não somente apoiador de dado partido, mas também seu representante nas urnas. Além disso, com o afastamento para exercício da atividade política, o servidor candidato também aumenta sua ligação com agentes políticos, podendo acarretar em maior influência política em suas atividades de interesse público, o que traz aumento do risco de corrupção inerente ao servidor. Assim, as informações de candidatura se mostram úteis à avaliação de corruptibilidade pretendida por este trabalho.

Por conseguinte, fazendo uso do apoio dos especialistas da DIE, a partir dos dados fornecidos pelo TSE foram selecionadas nove informações diferentes no âmbito de candidatura eleitoral, a serem futuramente processadas na fase de preparação dos dados para se tornarem um ou mais atributos. São elas:

1. **Valor máximo de despesa em campanha:** quantia máxima que detentor de candidatura, incluindo seu vice quando houver, pode gastar com despesas de campanha. Pode ser calculado diretamente a partir do valor numérico do campo valor de despesa máxima em campanha nos dados de candidatos do TSE.

2. **Ano da candidatura:** ano referente às eleições onde dado servidor público federal foi candidato a cargo eletivo. Pode ser calculado a partir do campo número de ano da eleição nos dados de candidatos do TSE.

3. **Número de turnos:** quantidade de turnos das eleições onde dado servidor público federal foi candidato a cargo eletivo. Pode ser calculado a partir do campo número de turno nos dados de candidatos do TSE.

4. **Cargo do candidato:** valor categórico da classificação do tipo de cargo eletivo pretendido pelo candidato servidor, com onze tipos diferentes nas amostras obtidas, como, por exemplo: Prefeito, Governador e Deputado Federal. Pode ser extraído do campo cargo do candidato nos dados de candidatos do TSE.

5. **Situação da candidatura:** valor categórico da descrição da situação da candidatura de cada servidor público federal, com seis tipos diferentes nas amostras obtidas, como, por exemplo: deferido, indeferido e renúncia. Pode ser extraído do campo situação da candidatura nos dados de candidatos do TSE.

6. **Partido da candidatura:** valor categórico da indicação da sigla do partido pelo qual dado servidor público federal se candidatou, com 36 tipos diferentes nas amos-

tras obtidas, como, por exemplo: PT, PSDB e DEM. Pode ser extraído do campo sigla do partido do candidato nos dados de candidatos do TSE.

7. **Grau de instrução do candidato:** valor categórico do grau de escolaridade de cada servidor público federal candidato, com 19 tipos diferentes nas amostras obtidas, como, por exemplo: superior incompleto, ensino médio e superior completo. Pode ser extraído do campo grau de instrução do candidato nos dados de candidatos do TSE.

8. **Estado civil do candidato:** valor categórico do estado civil de cada servidor público federal candidato, com nove tipos diferentes nas amostras obtidas, como, por exemplo: solteiro, casado e divorciado. Pode ser extraído do campo estado civil do candidato nos dados de candidatos do TSE.

9. **Situação do turno:** valor categórico da descrição da situação dos turnos das eleições da qual cada servidor público federal foi candidato, com 16 tipos diferentes nas amostras obtidas, como, por exemplo: eleito, suplente e não eleito. Pode ser extraído do campo situação do turno nos dados de candidatos do TSE.

3.2.4 Dimensão de Vínculos Societários

A dimensão de vínculos societários trata das informações de empresas que possuem relação de vínculo com servidores públicos federais. Esta dimensão compreende tanto informações dissociadas da Administração Pública, como dados cadastrais das empresas e doações a partidos políticos e comitês eleitorais, quanto dados resultantes da atuação em conjunto com o poder público, como recebimentos de recursos, sanções no âmbito público e terceirização na esfera federal. Assim, a dimensão de vínculos societários pode ter suas informações divididas em cinco grupos: dados cadastrais, impedimentos, doações políticas, recebimento de recurso público e terceirização.

3.2.4.1 Dados Cadastrais

Os dados cadastrais de empresas com relação de vínculo com servidores públicos federais abordadas pelo presente trabalho envolvem informações do cadastro de pessoa jurídica da Receita Federal e das relações empregatícias contidas na RAIS (Relação Anual de Informações Sociais).

Cadastro da RFB

A Secretaria da Receita Federal do Brasil (RFB) possui como uma de suas competências a administração do Cadastro Nacional da Pessoa Jurídica (CNPJ)[5]. O CNPJ compreende as informações cadastrais das entidades de interesse das administrações tributárias da União, dos Estados, do Distrito Federal e dos Municípios. Tal cadastro possui, dentre outras informações, a Ficha Cadastral da Pessoa Jurídica (FCPJ) e o Quadro de Sócios e Administradores (QSA) das empresas atuantes no Brasil.

Buscando descrever de maneira estruturada as áreas de atuação das empresas cadastradas no CNPJ, a RFB utiliza a Classificação Nacional de Atividades Econômicas (CNAE)[6]. Dessa forma, verifica-se que a CNAE é o instrumento de padronização nacional dos códigos de atividade econômica, podendo compreender estabelecimentos de empresas privadas ou públicas, organismos públicos e privados, instituições sem fins lucrativos e agentes autônomos. Tal classificação é estruturada de forma hierarquizada em cinco níveis, com 21 seções, 87 divisões, 285 grupos, 672 classes e 1318 subclasses. Com as combinações resultantes desses códigos foi possível identificar 4.913 códigos diferentes na base disponível para a DIE em maio de 2015. Atualmente já está em uso a CNAE 2.2, entretanto, as informações relativas a essa atualização ainda não estão disponíveis à CGU. É útil destacar ainda que cada empresa, para atingir seus objetivos, desenvolve uma atividade principal e pode desenvolver várias atividades secundárias, todas elas classificadas segundo a hierarquia da CNAE.

Com o objetivo de realizar a identificação da constituição jurídico-institucional das entidades públicas e privadas nos cadastros da Administração Pública do país, a RFB utiliza os códigos de Natureza Jurídica (NJ). No Brasil, as NJ são divididas em cinco grupos: Administração Pública; Entidades Empresariais; Entidades sem Fins Lucrativos; Pessoas Físicas; e Organizações Internacionais e Outras Instituições Extraterritoriais. Os códigos dessas naturezas possuem quatro dígitos e se iniciam pelo mesmo algarismo que representa o grupo. Em consulta às bases de dados da DIE sobre naturezas jurídicas, foram identificados 88 códigos diferentes. Ainda em relação às diversas naturezas jurídicas presentes no sistema CNPJ, cabe conferir um tratamento diferenciado no modelo às Empresas Públicas e Sociedades de Economia Mista, vez que o capital dessas entidades é total ou parcialmente público – Empresas Públicas e Sociedades de Economia Mista, respectivamente – logo é natural que as pessoas que compõem o Quadro de Sócios e Administradores sejam agentes públicos, assim, tais vínculos não devem ser considerados.

Além das classificações de atividades e natureza jurídica, a RFB também categoriza o porte das empresas cadastradas no CNPJ. São quatro classificações distintas em relação

[5]Cadastro Nacional da Pessoa Jurídica: CNPJ – Link: http://www.receita.fazenda.gov.br/PessoaJuridica/cnpj/ConsulSitCadastralCnpj.htm

[6]Classificação Nacional de Atividades Econômicas: CNAE – Link: http://www.receita.fazenda.gov.br/PessoaJuridica/CNAEFiscal/txtcnae.htm

ao porte: Microempresa (ME); Empresa de Pequeno Porte (EPP); Demais; e Sem Informação. Sendo que não foram encontradas nas bases disponíveis empresas com porte do tipo Sem Informação.

Outros fatos associados à importância da empresa remetem à situação do seu CNPJ, se ativa, suspensa, inapta, baixada ou nula. Essas situações são definidas pela Instrução Normativa RFB nº 1470 [19]. A inscrição no CNPJ é enquadrada na situação cadastral suspensa em função de diversas situações elencadas no art. 36 da referida Instrução Normativa. A condição de inapta ocorre em função dos artigos 42 e 43 da mesma norma, segundo os quais, dentre outras implicações, a pessoa jurídica cuja inscrição no CNPJ tenha sido declarada inapta é impedida de participar de concorrência pública, de celebrar convênios, acordos, ajustes ou contratos que envolvam recursos públicos. Já a situação baixada ocorre quando a entidade tiver sua solicitação de baixa deferida, ou tiver sua inscrição baixada de ofício. Enquanto que a inscrição nula é registrada quando for declarada a nulidade do ato de inscrição da entidade ou do estabelecimento filial, conforme art. 47 da Instrução Normativa indicada. E, finalmente, a inscrição no CNPJ é enquadrada na situação cadastral ativa quando a entidade não se enquadrar em nenhuma das demais situações, como consta no art. 35 da norma citada.

Jà em relação a qualificação do vínculo entre os servidores públicos federais e as empresas cadastradas no CNPJ, tem-se que dado servidor pode ser sócio, com diversas qualificações definidas, ou contador. Cada vínculo de sócio pode possuir associado a ele uma porcentagem de participação do sócio, que representa o grau de responsabilização na empresa vinculada. Além disso, dada pessoa física pode ser qualificada como responsável, que é aquele designado em estatuto, contrato social ou ata, incumbido de representar, ativa e passivamente, nos atos judiciais e extrajudiciais o agente regulado pessoa jurídica.

Portanto, vê-se que a gama de informações disponíveis no CNPJ pode ser de grande valia para a avaliação de risco de corrupção. Nesse sentido, considera-se que um servidor público ser sócio, responsável ou contador de entidade privada potencializa um conflito de interesses, em função da possibilidade desse servidor usar de influência direta – ou das condições, relacionamentos e oportunidades inerentes ao seu ambiente de trabalho – para obter vantagens de qualquer tipo para a empresa da qual participa, independentemente dessa vantagem trazer ou não prejuízo para a Administração Pública. Assim, dados do vínculo de um servidor público federal com uma empresa, assim como o porte e atividade econômica principal desta são informações que podem ter relação com o grau de corruptibilidade associado a um servidor.

Assim, seguindo o conhecimento dos especialistas da DIE, foram selecionadas 11 informações diferentes no âmbito do CNPJ, a serem futuramente processadas na fase de preparação dos dados para se tornarem um ou mais atributos. São elas:

1. **Quantidade de atividades secundárias:** número de atividades econômicas da CNAE tidas como secundárias da empresa com a qual cada servidor público federal possui vínculo. Pode ser calculado a partir do campo categórico atividade eliminando as denominadas principais pelo campo indicador principal nos dados de do CNPJ.

2. **Atividade principal:** valor categórico da atividade econômica da CNAE tida como principal da empresa com a qual cada servidor público federal possui vínculo. Como exemplo, pode-se citar: atividades de organizações associativas, construção de edifícios e obras de infraestrutura. Pode ser extraído a partir do campo categórico atividade filtrando as denominadas principais com o campo indicador principal nos dados de do CNPJ.

3. **Natureza jurídica:** valor categórico do tipo de natureza jurídica da empresa com a qual cada servidor público federal possui vínculo. Como exemplo, pode-se citar: Sociedade Anônima Fechada, Empresário Individual e Sociedade Simples Limitada. Pode ser calculado a partir do campo categórico natureza jurídica nos dados cadastrais de pessoas jurídicas do CNPJ.

4. **Grupo de natureza jurídica:** valor categórico do grupo de naturezas jurídicas da empresa com a qual cada servidor público federal possui vínculo. Como exemplo, pode-se citar: Entidades Empresariais, Entidades sem Fins Lucrativos e Pessoas Físicas. Pode ser calculado a partir do campo categórico grupo de natureza jurídica nos dados cadastrais de pessoas jurídicas do CNPJ.

5. **Situação da empresa:** valor categórico da descrição da situação da empresa com a qual cada servidor público federal possui vínculo. Como exemplo, pode-se citar: ativa, suspensa e nula. Pode ser extraído diretamente do campo situação nos dados cadastrais de pessoas jurídicas do CNPJ.

6. **Porte da empresa:** valor categórico do tipo de porte da empresa com a qual cada servidor público federal possui vínculo. Como exemplo, pode-se citar: Microempresas, Empresas de Pequeno Porte e demais. Pode ser calculado a partir do campo categórico porte da empresa nos dados cadastrais de pessoas jurídicas do CNPJ.

7. **Tempo de vínculo societário:** número de dias onde cada vínculo esteve ativo entre empresa e servidor público federal. Pode ser calculado a partir da diferença entre o campo data de entrada e data de saída nos dados de vínculos de sócios de pessoas jurídicas do CNPJ.

8. **Quantidade de vínculos vigentes:** número de vínculo ativos atualmente entre empresa e servidor público federal. Pode ser calculado a partir da quantidade de

valores diferentes de empresas pelo campo identificador único de CNPJ filtrando pelo campo data de saída vazia nos dados de vínculos de sócios de pessoas jurídicas do CNPJ.

9. **Porcentagem de participação do sócio:** valor numérico indicador da porcentagem de participação de cada servidor público federal na empresa com a qual possui vínculo. Pode ser extraído diretamente do campo porcentagem de participação nos dados de vínculos de sócios de pessoas jurídicas do CNPJ.

10. **Número de vínculos como responsável:** quantidade de vínculos com designação de responsável para dado servidor público federal em empresas onde possui vínculo. Pode ser calculado a partir da contabilização de registros do campo categórico indicador responsável nos dados de vínculos de sócios de pessoas jurídicas do CNPJ.

11. **Qualificação do vínculo:** valor categórico da descrição da qualificação de cada servidor público federal na empresa com a qual possui vínculo, aqui incluídos sócios e contadores. Pode-se citar como exemplo: Sócio-Administrador, Diretor e Presidente. Para sócios pode ser extraído do campo descrição de qualificação nos dados de vínculos de sócios de pessoas jurídicas do CNPJ.

Dados da RAIS

Os dados do Relatório Anual de Informações Sociais (RAIS) são instrumentos de coleta de dados instituídos pela Administração Pública, e tem por objetivo o suprimento às necessidades de controle da atividade trabalhista no país, o provimento de dados para a elaboração de estatísticas do trabalho e a disponibilização de informações do mercado de trabalho às entidades governamentais.

Assim, por meio da base de dados da RAIS disponível na DIE é possível verificar a quantidade de empregados que cada empresa cadastrou por ano, dimensionando as entidades em função do seu quantitativo assalariado. Outra informação também registrada na RAIS é a contabilização salarial de cada funcionário por empresa nos anos onde se tem dados.

No caso de empresas registradas com quantitativo de empregados igual a zero, vê-se situação clara de risco de corrupção, principalmente porque empresas sem empregados podem ter uma propensão maior a não entregar serviços ou produtos acordados. Dessa forma, as informações trabalhistas se mostram úteis na avaliação de corruptibilidade, ao possibilitar dimensionar empresas por funcionários e salário pago.

Com isso em mente, apoiando-se no conhecimento dos especialistas da DIE, duas informações diferentes foram definidas no âmbito da RAIS, a serem futuramente processadas na fase de preparação dos dados para se tornarem um ou mais atributos. São elas:

1. **Quantidade de funcionários:** número de funcionários da empresa com a qual dado servidor público federal possui vínculo. Pode ser calculado através da contabilização de identificadores únicos de pessoa física no campo funcionário para cada empresa inserida no campo empregador nos dados de vínculos empregatícios da RAIS.
2. **Salário:** valor do salário de funcionário pertencente a empresa com a qual dado servidor público federal possui vínculo. Pode ser calculado através do campo valor total de salário por empregado registrado no campo funcionário para cada empresa inserida no campo empregador nos dados de vínculos empregatícios da RAIS.

3.2.4.2 Impedimentos

Os dados de impedimentos no âmbito público de empresas com relação de vínculo com servidores públicos federais abordadas pelo presente trabalho envolvem informações de dois cadastros, a saber: Cadastro Nacional de Empresas Inidôneas e Suspensas (CEIS) e Cadastro de Entidades Privadas Sem Fins Lucrativos Impedidas (CEPIM).

CEIS

O Cadastro Nacional de Empresas Inidôneas e Suspensas (CEIS) é um banco de dados que tem por finalidade consolidar e divulgar a relação de empresas ou profissionais que sofreram sanções que tenham como efeito restrição ao direito de participar em licitações ou de celebrar contratos com a Administração Pública. Foi instituído pela Portaria CGU nº 516 [13], que desenvolveu e mantém o Sistema Integrado de Registro do CEIS[7], alimentado diretamente pelos entes da Administração Pública. É útil destacar que o artigo 23 da Lei nº 12.846 [18], denominada Lei Anticorrupção, trouxe a obrigatoriedade de os entes públicos, de todos os Poderes e Esferas de Governo, manterem esse cadastro atualizado.

De acordo com a Portaria citada, o CEIS conterá o registro das seguintes sanções:

1. suspensão temporária de participação em licitação e impedimento de contratar com a Administração;
2. declaração de inidoneidade para licitar ou contratar com a Administração Pública;

[7]Sistema Integrado de Registro do CEIS. – Link: http://www.cgu.gov.br/assuntos/responsabilizacao-de-empresas/sistema-integrado-de-registro-do-ceis-cnep

3. impedimento de licitar e contratar com a União, Estados, Distrito Federal ou Municípios;
4. proibição de contratar com o Poder Público e receber benefícios e incentivos;
5. proibição de participar de licitações e de contratar com o Poder Público;
6. declaração de inidoneidade pelo Tribunal de Contas da União; e
7. outras sanções previstas em legislações específicas ou correlatas com efeito de restrição ao direito de participar em licitações ou de celebrar contratos com a Administração Pública.

Vê-se pelo exposto que a inclusão de determinada empresa ou pessoa física no CEIS é a consequência de uma sanção aplicada por qualquer ente público, prevista no atual ordenamento jurídico em diversos dispositivos. Assim, o registro no CEIS pode estar relacionada a casos de corrupção, pois condutas como atos de improbidade, inexecução total ou parcial de contrato com a Administração e atividades lesivas ao meio ambiente ensejam sanções registradas no cadastro. Portanto, o fato de dado servidor público federal possuir vínculo societário com empresas que estão ou estiveram no CEIS é relevante e merece ser considerado na avaliação do risco de corrupção.

Nesse sentido, a partir do direcionamento dado pelos especialistas da DIE, três informações diferentes foram definidas no que se refere ao CEIS, a serem futuramente processadas na fase de preparação dos dados para se tornarem um ou mais atributos. São elas:

1. **Tipo de sanção:** valor categórico com descrição do tipo da sanção imputada à empresa com a qual cada servidor público federal possui vínculo. Como exemplo, pode-se citar: proibição, impedimento e suspensão. Pode ser extraído diretamente do campo categórico tipo de sanção nos dados do CEIS.
2. **Fundamentação legal:** valor categórico com fundamentação legal da sanção imputada à empresa com a qual cada servidor público federal possui vínculo. Como exemplo, pode-se citar: Lei nº 8.429/1992 - Lei de Improbidade [5] e Lei nº 8.443/1992 - Lei Orgânica TCU [6]. Pode ser extraído diretamente do campo categórico fundamentação legal nos dados do CEIS.
3. **Tempo de sanção:** quantidade de meses onde cada sanção esteve vigente no CEIS para empresa vinculada a servidor público federal. Pode ser calculado a partir da diferença entre o campo data de início de sanção e data de término de sanção nos dados do CEIS.

Das informações acima, vê-se problemas de qualidade nos campos atinentes à informação de tempo de sanção, a saber: data de início de sanção e data de término de sanção. Torna-se necessária, portanto, uma limpeza dos dados de tais campos.

CEPIM

O Cadastro de Entidades Privadas Sem Fins Lucrativos Impedidas (CEPIM)[8] é um banco de informações mantido pela CGU, a partir de dados fornecidos pelos órgãos e entidades da Administração Pública Federal, que tem por objetivo consolidar e divulgar a relação das entidades privadas sem fins lucrativos que estão impedidas de celebrar convênios, contratos de repasse ou termos de parceria com a Administração Pública Federal, nos termos do Decreto n.º 7.592 [14]. No atual ordenamento jurídico, vários são os dispositivos legais que restringem o direito de entidades privadas sem fins lucrativos celebrarem convênios, contratos de repasse ou termos de parceria com a Administração Pública Federal e de receberem transferências de recursos, participarem de licitações ou de celebrar contratos com a Administração Pública.

Os registros constantes do CEPIM tem como base as informações inseridas no Sistema Integrado de Administração Financeira (SIAFI)[9] pelos órgãos e entidades da Administração Pública Federal concedentes de recursos. Tais registros são disponibilizados na base de dados do CEPIM no dia seguinte a sua inserção. São consideradas impedidas as entidades privadas sem fins lucrativos que estejam registradas no SIAFI em "Inadimplência Efetiva" ou "Impugnados". Assim, apesar da DIE possuir acesso aos dados históricos, os impedimentos ficam vigentes apenas enquanto durarem as sanções correspondentes.

Nesse contexto, a existência de uma entidade cadastrada no CEPIM já implica na ocorrência de irregularidades por parte da mesma, de modo que quaisquer servidores públicos federais vinculados a esta entidade podem ter exercido influência em tais resultados, aumentando o risco de corrupção associado. Além disso, o próprio fato de haver vínculo entre um servidor e uma entidade que firme convênios com a Administração Pública já pode favorecer uma situação de conflito de interesses. Portanto, pelo exposto, vê-se a necessidade de utilizar os dados cadastrais do CEPIM na avaliação de corruptibilidade.

Por conseguinte, com o auxílio dos especialistas da DIE, foram selecionadas três informações diferentes para o CEPIM, a serem futuramente processadas na fase de preparação dos dados para se tornarem um ou mais atributos. São elas:

1. **Quantidade de impedimentos:** número de impedimentos, vigentes ou não, de entidades com as quais cada servidor público federal possui vínculo. Pode ser cal-

[8] Cadastro de Entidades Privadas Sem Fins Lucrativos Impedidas (CEPIM). – Link: http://www.portaltransparencia.gov.br/cepim/

[9] Site principal do SIAFI: http://www.tesouro.fazenda.gov.br/siafi

culado a partir da contabilização de registros onde cada entidade identificada pelo campo CNPJ aparece nos dados do CEPIM.

2. **Quantidade de impedimento vigentes:** número de impedimentos vigentes de entidades com as quais cada servidor público federal possui vínculo. Pode ser calculado a partir da contabilização de registros onde cada entidade identificada pelo campo CNPJ aparece nos dados do CEPIM, filtrando-se pelo campo data de exclusão igual a nulo.

3. **Tempo de impedimento:** quantidade de dias onde cada impedimento esteve vigente no CEPIM para entidade vinculada a servidor público federal. Pode ser calculado a partir da diferença entre o campo data de inclusão e data de exclusão nos dados do CEPIM.

3.2.4.3 Doações Eleitorais

Segundo o art. 39 da Lei nº 9.096 [7], partidos políticos podem receber doações de pessoas físicas e jurídicas para constituição de seus fundos, sendo que para arrecadar e aplicar recursos, os partidos e candidatos podem o fazer diretamente através da obtenção de CNPJ ou podem ser registrados Comitês Financeiros. Tem-se pela Lei que especificamente pessoas jurídicas poderão doar até 2% (dois por cento) do seu faturamento bruto no ano anterior ao da eleição. Além disso, o art. 36 da referida Lei define também que no caso de recebimento de doações cujo valor ultrapasse os limites previstos, será aplicada ao partido multa correspondente ao valor que exceder aos limites fixados. É útil observar ainda que as doações efetuadas a partidos políticos não são dedutíveis nas declarações de impostos de pessoa jurídica. Dessa forma, a doação não traz diretamente benesse alguma a uma pessoa jurídica que a realize, podendo inclusive acarretar em multa caso a doação exceda limites definidos em lei. Assim, a doação deve-se praticamente estritamente a um apoio da empresa ao candidato ou partido, o que implica a proximidade no mínimo ideológica. Consequentemente, servidores públicos federais pertencentes ao quadro societário de tais empresas também podem se identificar com a ideologia partidária apoiada pela doação.

Tendo isso em vista e frente ao já exposto na Seção 3.2.3.1, vê-se que a relação de proximidade entre servidores e partidos políticos pode acarretar a consecução de atividades públicas com direcionamento político em contraste ao interesse público. Assim, considerou-se válido avaliar neste trabalho as informações de doação a partidos, candidatos e comitês, de empresas vinculadas a servidores públicos federais, buscando mensurar o grau de proximidade entre as partes envolvidas.

Nesse sentido, com o intuito de contribuir com a transparência do processo eleitoral brasileiro, o Tribunal Superior Eleitoral (TSE) divulga em sua página de repositório de dados eleitorais[10] diversas informações relativas a prestação de contas – incluindo receitas e despesas de campanha dos candidatos, partidos e comitês – atualmente disponíveis em meio digital para as eleições de 2002 a 2014.

Com isso, através da consolidação dos dados obtidos do repositório do TSE, foram criadas bases de dados na DIE com as informações de cada eleição disponível. Separando-se por ano, tem-se discriminado, em tabelas diferentes, informações relativas a receitas de candidatos, comitês e partidos. Com a identificação do CNPJ do doador é possível, então, contabilizar o montante doado registrado como receita.

Dessa forma, através da orientação dos especialistas da DIE, foram extraídas duas informações diferentes relacionadas a doações de empresas a partidos, comitês ou candidatos, que serão processadas na fase de preparação dos dados para se tornarem um ou mais atributos. São elas:

1. **Número de empresas que realizaram doação:** quantidade de empresas diferentes, vinculadas a servidores públicos federais, que realizaram doações a partidos, comitês ou candidatos. Pode ser calculada a partir da soma do número de diferentes valores do campo de identificação única de empresas que efetuaram doações em cada tabela de receita por ano.

2. **Montante doado pelas empresas:** valor doado a partidos, comitês ou candidatos por empresas vinculadas a servidores públicos federais. Pode ser calculado através da soma de valores registrados como receita a cada ano por empresa doadora em cada tabela de receita por ano.

Das informações acima, vê-se problemas de qualidade nos campos atinentes à informação de montante doado, em praticamente todos os anos onde há dados. Torna-se necessária, portanto, uma limpeza dos dados de tal campo.

3.2.4.4 Recebimento de Recurso Público

Os dados de recebimentos de recursos públicos por parte de empresas com relação de vínculo com servidores públicos federais abordadas pelo presente trabalho envolvem informações de convênios cadastrados no Sistema de Gestão de Convênios (SICONV)[11] e Ordens Bancárias (OB) registradas no Sistema Integrado de Administração Financeira do Governo Federal (SIAFI)[12].

[10]Repositório de dados eleitorais do TSE: http://www.tse.jus.br/hotSites/pesquisas-eleitorais/prestacao_contas.html

[11]Página do Sistema de Convênios (SICONV): https://www.convenios.gov.br/portal/

[12]Site principal do SIAFI: http://www.tesouro.fazenda.gov.br/siafi

SICONV

Segundo o Decreto nº 6.170 [11], Convênios são acordos ou outros instrumentos que disciplinam a transferência de recursos financeiros da esfera federal para órgão ou entidade da Administração Pública Federal, direta ou indireta, ou entidades privadas sem fins lucrativos. Possui como objetivo viabilizar a aplicação de recursos em determinadas áreas de interesse da sociedade, buscando atingir uma maior efetividade das políticas públicas. Pressupõe uma convergência de propósitos entre duas partes, que celebram um acordo visando interesses em comum.

Nos convênios, por um lado, tem-se como Concedente o órgão ou entidade da Administração Pública que, por meio da celebração de convênio, repassa recursos da União a ente público ou privado. Já Convenente é o ente, público ou privado, beneficiário de recursos da União repassados pelo concedente em função da celebração de convênio. No âmbito da realização de convênios, há ainda a figura da empresa ou entidade subcontratada, que é aquela pessoa jurídica que firma contrato oneroso com a entidade convenente para fornecimento de bens e/ou serviços necessários à consecução do objeto do convênio.

O Sistema de Gestão de Convênios (SICONV), oficialmente implantado em 2008, foi concebido para aumentar a transparência do gasto público federal realizado pelo instrumento. Nos anos após sua implantação, inúmeros escândalos de corrupção foram detectados[13] [14] [15] por meio de simples consultas e cruzamentos de dados no SICONV – nesse sentido, a própria DIE trabalhou com cruzamento de dados e obteve diversos resultados já publicados [59].

Na análise de ocorrência de corrupção, tem-se que o cerne da manifestação em convênios reside na discricionariedade atribuída ao órgão concedente – em última análise, a um servidor público – na escolha da entidade que irá executar determinado convênio. A legislação, em nome dos princípios da isonomia, impessoalidade, moralidade e eficiência, obriga o concedente a proceder a um chamamento público para seleção da entidade mais apta a executar o convênio, conforme art. 4º do Decreto 6.170 [11]. No entanto, na prática, a escolha muitas vezes dá-se em função de relações pessoais ou em atendimento a pleitos de agentes políticos, situações que retiram do instrumento de transferências vo-

[13] Notícia de 12 de setembro de 2013 da Operação Esopo, sobre fraude em convênios de prestação de serviços: http://noticias.uol.com.br/politica/ultimas-noticias/2013/09/12/instituto-privado-e-suspeito-de-causar-prejuizo-de-r-26-mi-em-programa-do-mte.htm

[14] Notícia de 16 de janeiro de 2014 da Operação Pronto Emprego, a respeito de fraudes em convênios com Ministério do Trabalho: http://memoria.ebc.com.br/agenciabrasil/noticia/2014-01-16/mpf-denuncia-14-pessoas-por-fraudes-em-convenios-com-ministerio-do-trabalho

[15] Notícia de 1 de novembro de 2012 sobre detecção de fraudes em convênios do Ministério da Cultura: http://politica.estadao.com.br/noticias/geral,tcu-detecta-fraude-em-convenios-do-minc-com-28-ongs,954518

luntárias o caráter de "acordo de vontades" para realização do bem coletivo, para em seu lugar prevalecer, em muitos casos, a submissão a interesses privados ilegítimos.

Dessa forma, observa-se que as informações de convênios se mostram úteis na avaliação de corruptibilidade, principalmente no que tange a vínculos de servidores públicos federais com entidades convenentes ou empresas subcontratadas em convênios federais, devido ao possível conflito de interesses.

Nesse sentido, a partir do esclarecimento dos especialistas da DIE, quatro informações diferentes foram definidas para o estudo da corruptibilidade relacionada com vínculos de servidores com entidades e/ou empresas participantes de convênios. Tais dados serão futuramente processados na fase de preparação dos dados para se tornarem um ou mais atributos, a saber:

1. **Número de entidades ou empresas que participaram de convênio:** quantidade de entidades ou empresas diferentes, vinculadas a servidores públicos federais, que participaram de convênios, sejam como convenentes ou subcontratadas. Pode ser calculada a partir da soma do número de diferentes valores do campo de identificação única de empresas que participaram de convênios, tanto como convenentes ou subcontratadas, nas tabelas de partícipe e fornecedores, respectivamente.

2. **Número de convênios:** quantidade de convênios diferentes com participação de entidades ou empresas, como convenentes ou subcontratadas, vinculadas a servidores públicos federais. Pode ser calculada a partir da soma do número de valores diferentes do campo sequencial de convênios, nas tabelas de partícipe e fornecedores.

3. **Número de empresas subcontratadas em convênios:** quantidade de empresas diferentes, vinculadas a servidores públicos federais, que participaram de convênios como subcontratadas. Pode ser calculada a partir da soma do número de diferentes valores do campo de identificação única de empresas que participaram de convênios como subcontratadas, na tabela fornecedores.

4. **Número de entidades convenentes em convênios:** quantidade de entidades diferentes, vinculadas a servidores públicos federais, que participaram de convênios como convenentes. Pode ser calculado a partir da soma do número de diferentes valores do campo de identificação única de empresas que participaram de convênios como convenentes, na tabela partícipe.

OB

No governo federal o pagamento de despesas é realizado por meio do Sistema Integrado de Administração Financeira do Governo Federal (SIAFI), com a emissão de Ordem Ban-

cária (OB)[16], documento que possui várias espécies e características próprias, variando de acordo com o tipo de pagamento a ser realizado.

Conforme Manual SIAFI – Macrofunções, Assunto 020305 - Conta Única do Tesouro Nacional[17] – as ordens bancárias são classificadas em 14 tipos. Considerando que a finalidade das informações em análise é assinalar potenciais riscos de corrupção derivados de vínculos societários de servidores públicos com empresas fornecedoras, não cabe utilizar no modelo todos os tipos de OB disponíveis, vez que algumas não representam pagamentos a fornecedores ou prestadores de serviço. Várias OB listadas na tabela importam em meros desembolsos derivados de rotinas administrativas ou judiciais, com pouca margem para discricionariedade do dirigente público. Nesse sentindo, as ordens bancárias identificadas como pagamentos compatíveis para integrar a análise de materialidade são:

1. Ordem Bancária de Crédito (OBC) – códigos 11, 12, 14 e 21 no SIAFI: utilizada para pagamentos por meio de crédito em conta-corrente do favorecido na rede bancária. O favorecido deverá ser obrigatória e exclusivamente um CPF ou CNPJ. É o tipo mais comum de Ordem Bancária.

2. Ordem Bancária de Banco (OBB) – código 13 no SIAFI: utilizada para pagamento de documentos em que o agente financeiro deva dar quitação e que não seja possível o pagamento por OB fatura, bem como para contratação de câmbio com outros bancos que não o Banco do Brasil. Para pagamentos a diversos credores ou para folha de pessoal deve ser anexada à OBB uma lista de credores.

3. Ordem Bancária de Fatura (OBD) – código 59 no SIAFI: utilizada para pagamento de título de cobrança/boletos bancários, pela unidade gestora, com uso de código de barras. Como exemplo, têm-se os boletos emitidos para pagamento de fatura de concessionárias de água, energia e telefone ou para quitação de tributos estaduais (IPVA) e municipais (ISS), junto aos respectivos governos.

Considerando-se que a existência de pagamentos, via ordens bancárias, a fornecedores de cujos quadros societários participe servidor público federal é capaz de configurar um potencial conflito entre os interesses público e particular, tem-se que as informações citadas podem ser úteis na avaliação de corrupção. Assim, através da orientação dos especialistas da DIE, foram extraídas três informações diferentes, que serão processadas na fase de preparação dos dados para se tornarem um ou mais atributos. São elas:

[16]Manual Simplificado de Ordens Bancárias: http://www.tesouro.fazenda.gov.br/in/manual-ordem-bancaria

[17]Manual SIAFI: Macrofunções, Assunto 020305 - Conta Única do Tesouro Nacional: http://manualsiafi.tesouro.fazenda.gov.br/020000/020300/020305/

1. **Número de empresas que receberam OB:** quantidade de empresas diferentes, vinculadas a servidores públicos federais, que receberam OBC, OBB ou OBD. Pode ser calculada a partir da soma do número de diferentes valores do campo de identificação única de empresas que receberam OB – com filtro de tipo de OB para OBB, OBC e OBD – na tabela de ordens bancárias.

2. **Número de Ordens Bancárias recebidas:** quantidade de Ordens Bancárias – OBC, OBB ou OBD – recebidas por empresas vinculadas a servidores públicos federais. Pode ser calculada a partir da soma do número de diferentes valores do campo de identificação de número de ordem bancária – com filtro de tipo de OB para OBB, OBC e OBD – na tabela de ordens bancárias.

3. **Valor de OB recebida:** valor de OB – OBC, OBB ou OBD – recebida por empresas vinculadas a servidores públicos federais. Pode ser calculado a partir da soma dos valores recebidos para cada valor do campo de identificação única de empresas que receberam OB – com filtro de tipo de OB para OBB, OBC e OBD – na tabela de Ordens Bancárias.

3.2.4.5 Terceirização

A partir da década de 60, com o Decreto-Lei nº 200 [2], a terceirização no âmbito da Administração Pública começou a ser disciplinada, seguindo a orientação, conforme art. 10, de que a execução das atividades da Administração Federal deverá ser amplamente descentralizada. Segundo art. 1º do Decreto nº 2.271 [8], no âmbito da Administração Pública Federal direta, autárquica e fundacional, poderão ser objeto de execução indireta as atividades materiais acessórias, instrumentais ou complementares aos assuntos que constituem área de competência legal do órgão ou entidade. Como exemplo de objetos terceirizáveis, tem-se as atividades de conservação, limpeza, segurança, vigilância, transportes, informática, copeiragem, recepção, reprografia, telecomunicações e manutenção de prédios, equipamentos e instalações.

Conforme o Ofício-circular nº 268/2009/SE/CGU-PR [12], a cada quatro meses, os ministérios do Poder Executivo federal, incluindo Autarquias e Fundações, devem enviar à CGU informações acerca dos contratos de terceirização de mão-de-obra. Assim, através do Sistema de Transferência de Informações (STI)[18] da CGU os dados de contratos cujos terceirizados são contratados com dedicação exclusiva de mão-de-obra são cadastrados. Útil observar a especificidade de dedicação exclusiva, pois contratados por regime de empreitada por preço unitário (contratação por produto) ou contratos classificados na

[18]Sistema de Transferência de Informações (STI) da CGU: https://sti.cgu.gov.br/

natureza de despesa "Outros Serviços de Terceiros" ou regime de empreitada por preço unitário, não são informados ao STI.

Com as informações registradas pelas unidades do Poder Executivo federal, têm-se dados diversos de terceirizados desde o ano de 2010, incluindo valores mensais de salário e custo (salário com a inclusão de encargos trabalhistas). A existência de empresas vinculadas a servidor público que possuem funcionários como terceirizados na Administração Pública pode ser vista como situação de risco, dentre outros aspectos, a medida que os terceirizados sejam empregados devido ao relacionamento entre a empresa e um dado servidor público, como troca de favores ou outro acordo. Assim, de modo a viabilizar a mensuração da relação entre terceirizados e servidores públicos federais com vistas a avaliar corruptibilidade, este trabalho considera válido incluir dados de empresas que possuem vínculo com servidores e empregam terceirizados na Administração Pública Federal.

Nesse sentido, a partir do direcionamento dado pelos especialistas da DIE, cinco informações diferentes foram definidas no que se refere a terceirização, a serem futuramente processadas na fase de preparação dos dados para se tornarem um ou mais atributos. São elas:

1. **Quantidade de empresas com terceirização:** quantidade de empresas diferentes, vinculadas a servidores públicos federais, que possuem funcionários empregados como terceirizados na Administração Pública Federal. Pode ser calculada a partir da soma do número de diferentes valores do campo de identificação única de empresas que possuem terceirizados na tabela de terceirizados.

2. **Quantidade de unidades com terceirizados:** quantidade de unidades da Administração Pública Federal diferentes com empregados terceirizados de empresas vinculadas a servidores públicos federais. Pode ser calculada a partir da soma do número de diferentes valores do campo de código de unidade gestora com terceirizados de empresas na tabela de terceirizados.

3. **Valor mensal de salário de terceirizado:** quantia de salário que terceirizados empregados de empresas vinculadas a servidores públicos federais recebem mensalmente. Pode ser calculado a partir do campo de valor mensal de salário de empresas na tabela de terceirizados.

4. **Valor mensal de custo de terceirizado:** quantia de salário mais encargos trabalhistas que terceirizados empregados de empresas vinculadas a servidores públicos federais recebem mensalmente. Pode ser calculada a partir do campo de valor mensal de custo de empresas na tabela de terceirizados.

5. **Número de anos diferentes com terceirização:** quantidade de anos diferentes havendo terceirizados empregados de empresas vinculadas a servidores públicos federais. Pode ser calculado a partir do número de valores distintos do campo de número do ano de empresas na tabela de terceirizados.

3.3 Preparação dos Dados

A preparação dos dados consiste no processamento e ajuste das diversas informações levantadas nos entendimentos do negócio e dos dados. Esta fase começa com a limpeza de inconsistências nos dados e passa pela construção dos atributos a serem utilizados na modelagem. Além do tratamento de valores faltantes e transformação de atributos categóricos em numéricos, tem-se a análise da variância e correlação dos atributos resultantes, etapas comumente feitas em seleção de atributos [31], de forma a observar a necessidade de eliminação de alguns deles.

Os dados a serem preparados são extraídos para duas classes, chamadas: "Corruptos" e "Não Corruptos". Os servidores corruptos são aqueles pertencentes ao CEAF que foram expulsos por motivo de corrupção, conforme definido na Seção 3.2.1. Os não corruptos na verdade são de corruptibilidade desconhecida, pois é impossível atestar tal característica. Assim, considerando que os corruptos são minoria dentro de um grupo maior de servidores, serão realizadas amostragens no grupo de não corruptos – ou de corruptibilidade desconhecida – nas etapas de modelagem, a serem vistas na Seção 3.4, de forma a minimizar o impacto da possibilidade de haver corruptos ainda não expulsos no conjunto de dados de não corruptos. No entanto, é importante perceber que é esperado que alguns dos servidores no grupo de não corruptos sejam classificados como corruptos, pois devido ao desconhecimento da corruptibilidade real de tal grupo, há de fato nele alguns corruptos.

Assim, considerando que serão feitas diversas amostragens, o conjunto de dados para os servidores não corruptos foi criado a partir de uma amostra aleatória de aproximadamente 300.000 servidores públicos federais com CPF válido – quantidade aproximadamente 60 vezes maior que a de corruptos – seguida de um filtro removendo aqueles cadastrados no CEAF, número que permite as reamostragens necessárias e possibilita a manipulação dos dados na ferramenta estatística R através da IDE RStudio[19] com o equipamento disponível para este trabalho.

Tem-se ainda na preparação dos dados, a separação de dados inicial realizada nos conjuntos de corruptos e não corruptos, descrita na Seção 3.3.4, seguindo-se para as atividades de ajuste de atributos de forma a subsidiar a seleção dos mais relevantes na

[19]RStudio – IDE para ferramenta estatística R. Link: http://www.rstudio.com/

Seção 3.3.5, atividade esta feita em um processo que envolve modelagem, descrito na Seção 3.4.1.

3.3.1 Limpeza de Dados

A etapa de preparação de dados denominada limpeza de dados envolve diversas tarefas com o objetivo de manter os dados consistentes e livre de erros, utilizando como principal direcionamento a indicação de problemas no Entendimento dos Dados, conforme apresentado na Seção 3.2. Como atividades de limpeza realizadas, pode-se citar:

- integração de dados de fontes distintas;
- padronização de valores descritos de maneira distinta mas de significados equivalentes;
- eliminação ou ajuste de dados individualmente inconsistentes;
- eliminação ou ajuste de dados que em conjunto se mostram inconsistentes;
- conversão de tipos de dados para formatos adequados; e
- padronização de tipos de dados.

Além das atividades mencionadas, o tratamento de valores faltantes também foi realizado. Para atributos categóricos foi criada uma categoria "NA" representando a inexistência de valores para dado atributo. Já para atributos relacionados a contagens, com valores decimais, os registros faltantes representam o próprio valor igual a zero, logo foram substituídos por tais valores. Além disso, outros campos foram tratados individualmente quanto a valores faltantes, como, por exemplo, data de cancelamento de filiação e data de saída de vínculo societário para servidores respectivamente com filiação ainda ativa e vínculo societário vigente – nesses casos, para cálculos de tempo foram usados com data final valores recentes, como, por exemplo, "20150501".

3.3.2 Construção de Atributos

Nesta seção, a partir dos dados selecionados no entendimento dos dados e de outros tidos como relevantes pelos especialistas da DIE, os atributos que efetivamente serão utilizados nas etapas seguintes são definidos em linhas gerais e seus passos de construção são brevemente comentados, separando as informações segundo a divisão de dimensões delineada no Entendimento dos Dados, conforme apresentado na Seção 3.2.

É válido atentar que alguns atributos categóricos serão criados como variáveis de contagem. Em tal transformação, para dado atributo com c categorias serão criadas c variáveis

binárias, uma para cada categoria indicando a presença (valor 1) ou ausência (valor 0) daquela categoria do atributo, de modo semelhante ao procedimento realizado para *dummy variables* [62]. Em seguida, como um mesmo servidor pode ter vários registros para um dado atributo, os valores das variáveis binárias de cada categoria são agrupados através de uma soma, gerando as variáveis de contagem. Dessa forma, os atributos que passarem por este processo nas seções seguintes serão indicados como variáveis de contagem.

3.3.2.1 Dimensão de Corrupção

Na dimensão de corrupção, a única informação utilizada é o próprio CPF de cada indivíduo expulso por motivos de corrupção. Na base de dados do CEAF há alguns CPFs com erro de inserção, sendo que através de informações de nome e órgão foi possível realizar a correção manual de alguns deles.

3.3.2.2 Dimensão Funcional

A partir dos dados básicos da Dimensão Funcional, fazendo uso principalmente das informações elencadas no Entendimento dos Dados, explicitadas na Seção 3.2.2.1, os seguintes atributos foram definidos, com *alias* em negrito:

- **siape.sal.bruto:** salário bruto de cada servidor. Como um mesmo servidor pode possuir ou já ter possuído mais de um salário, este atributo foi desdobrado em três, a saber: siape.sal.bruto.max, siape.sal.bruto.min e siape.sal.bruto.med;
- **siape.qtd.cargos:** quantidade de cargos diferentes que um dado servidor possui ou já possuiu;
- **siape.qtd.orgaos.cargo:** quantidade de órgãos diferentes em que um servidor está ou já esteve lotado possuindo cargo;
- **siape.qtd.orgaos.funcao:** quantidade de órgãos diferentes em que um servidor está ou já esteve lotado possuindo função;
- **siape.qtd.atividades:** quantidade de atividades diferentes que um servidor exerce ou já exerceu;
- **siape.dias.cargo:** tempo (em dias) em que um servidor esteve em dado cargo ou, para vínculos ainda ativos, tempo desde a entrada até uma data recente. Como um mesmo servidor pode estar ou já ter estado em mais de um cargo, este atributo foi desdobrado em quatro, a saber: siape.dias.cargo.total, siape.dias.cargo.med, siape.dias.cargo.max e siape.dias.cargo.min;

- **siape.situacao.funcional:** valor categórico da descrição de situação funcional do servidor. Como um mesmo servidor pode possuir ou já ter possuído mais de uma situação funcional, este atributo categórico foi transformado em variáveis de contagem, uma para cada categoria; e
- **siape.nivel.funcao:** valor categórico de nível de função do servidor. Como um mesmo servidor pode possuir ou já ter possuído mais de um nível de função, este atributo categórico foi transformado em variáveis de contagem, uma para cada categoria.

Para as informações de sanções, foram utilizadas as indicações da Seção 3.2.2.2 para criação dos atributos definidos a seguir, com *alias* explicitado em negrito:

- **tcuirregulares.qtd.total.contas.julg.irreg:** quantidade total de contas julgadas irregulares de um servidor;
- **tcuirregulares.origem.de.recursos:** valor categórico da origem dos recursos envolvidos no julgamento de contas de um servidor. Como um mesmo servidor pode possuir mais de uma origem de recursos em mais de uma conta julgada irregular, este atributo categórico foi transformado em variáveis de contagem, uma para cada categoria;
- **ativa.qtd.cert.ressalva:** número de certificados de regularidade com ressalva onde um servidor foi elencado como gestor;
- **ativa.qtd.cert.irregular:** número de certificados de irregularidade onde um servidor foi elencado como gestor;
- **ativa.qtd.const:** quantidade de constatações onde um servidor foi elencado como responsável;
- **ativa.qtd.os:** quantidade de OS envolvendo um servidor;
- **ativa.qtd.const.por.os:** quantidade de constatações por OS envolvendo um servidor;
- **ativa.tipo.const:** valor categórico da classificação da constatação onde um servidor foi elencado como responsável. Como um mesmo servidor pode possuir mais de um tipo constatação, este atributo categórico foi transformado em variáveis de contagem, uma para cada categoria; e
- **ativa.tipo.const.por.os:** a partir dos atributos quantitativos gerados pelo atributo ativa.ds.tipo.const, dividiu-se cada variável de contagem pela quantidade de OS, de forma a relativizar as quantidades.

Já em relação aos dados levantados em investigações descritos na Seção 3.2.2.3, os seguintes atributos foram definidos, com *alias* explicitado em negrito:

- **pad.qtd.penalidade.advertencia:** quantidade de procedimentos disciplinares onde um servidor recebeu a penalidade de advertência;
- **pad.qtd.penalidade.suspensao:** quantidade de procedimentos disciplinares onde um servidor recebeu a penalidade de suspensão;
- **pad.qtd.penalidade.susp.conv.em.multa:** quantidade de procedimentos disciplinares onde um servidor recebeu a penalidade de suspensão convertida em multa;
- **cgie.qtd.tarefas:** quantidade de procedimentos de investigação realizados pela CGIE onde um servidor consta como envolvido; e
- **cgie.tipo.envolvimento:** valor categórico do tipo de envolvimento de um servidor em uma investigação da CGIE. Como um mesmo servidor pode estar envolvido ou já sido envolvido em mais de uma investigação da CGIE, este atributo categórico foi transformado em variáveis de contagem, uma para cada categoria.

3.3.2.3 Dimensão Política

Após o entendimento dos dados da Dimensão Política, conforme Seção 3.2.3, utilizando as informações elencadas para filiação partidária nesta dimensão, os seguintes atributos foram definidos, com *alias* definido em negrito:

- **filiados.qtd.filiacao.SET:** quantidade de filiações realizadas no mês de setembro de um servidor. Este atributo foi elencado buscando levantar casos de filiações realizadas muito próximas do limite da data de filiação do TSE, geralmente fixada no mês de outubro, por poderem indicar casos de filiação não espontânea, para fins de interesses partidários;
- **filiados.qtd.filiacao.OUT:** quantidade de filiações realizadas no mês de outubro de um servidor. Este atributo foi elencado pelo mesmo motivo do anterior;
- **filiados.qtd.cancelamentos:** quantidade de cancelamentos de filiação de um servidor;
- **filiados.qtd.desfiliacoes:** quantidade de desfiliações de um servidor;
- **filiados.dias.filiacao:** tempo (em dias) em que um servidor esteve filiado ou, para filiações ainda ativas, tempo desde a filiação até uma data recente. Como um mesmo servidor pode estar filiado ou já ter sido filiado mais de uma vez, este atributo foi

desdobrado em quatro, a saber: filiados.dias.filiacao.total, filiados.dias.filiacao.med, filiados.dias.filiacao.max e filiados.dias.filiacao.min;

- **filiados.ds.motivo.cancelamento:** valor categórico da descrição do motivo de cancelamento de filiação de um servidor. Como um mesmo servidor pode ter cancelado mais de uma filiação, este atributo categórico foi transformado em variáveis de contagem, uma para cada categoria;
- **filiados.sg.partido:** valor categórico de sigla de partido ao qual um servidor se filia. Como um mesmo servidor pode estar filiado ou já ter se filiado a mais de um partido, este atributo categórico foi transformado em variáveis de contagem, uma para cada categoria; e
- **filiados.qtd.total.filiacoes:** quantidade total de ocorrências de filiação a partidos registradas para um servidor.

Já para os dados políticos de candidaturas eleitorais, conforme entendimento descrito na Seção 3.2.3.2, os atributos foram construídos como mostrado a seguir, com *alias* explicitado em negrito:

- **candidatos.qtd.cand.uf.eleic.igual.nasc:** quantidade de candidaturas onde um servidor foi candidato pela mesma UF (Unidade Federativa) onde nasceu;
- **candidatos.soma.vl.despesa.max.campanha:** soma dos valores máximos de despesa em campanha de um servidor candidato;
- **candidatos.nr.ano.eleicao:** ano referente às eleições onde um servidor foi candidato a cargo eletivo. Como um mesmo servidor pode ser candidato em mais de um ano, este atributo foi transformado em variáveis de contagem, uma para cada categoria;
- **candidatos.nr.turno:** turno referente às eleições onde um servidor foi candidato a cargo eletivo. Como um mesmo servidor pode ser candidato e participar de mais um turno em mais de uma eleição, este atributo foi transformado em variáveis de contagem, uma para cada categoria;
- **candidatos.cargo:** valor categórico da classificação do cargo eletivo pretendido por um candidato servidor. Como um mesmo servidor pode ser candidato em mais de um ano, este atributo foi transformado em variáveis de contagem, uma para cada categoria;
- **candidatos.sit.candidatura:** valor categórico da descrição da situação da candidatura de um servidor. Como um mesmo servidor pode ser candidato em mais de

um ano, este atributo foi transformado em variáveis de contagem, uma para cada categoria;

- **candidatos.cand.sg.partido:** valor categórico da indicação da sigla do partido pelo qual um servidor se candidatou. Como um mesmo servidor pode ser candidato em mais de um ano, este atributo foi transformado em variáveis de contagem, uma para cada categoria;
- **candidatos.cand.grau.instrucao:** valor categórico do grau de escolaridade de um servidor candidato. Como um mesmo servidor pode ser candidato em mais de um ano e seu grau de instrução pode variar em tais anos, este atributo foi transformado em variáveis de contagem, uma para cada categoria;
- **candidatos.cand.estado.civil:** valor categórico do estado civil de um servidor candidato. Como um mesmo servidor pode ser candidato em mais de um ano e seu estado civil variar em tais anos, este atributo foi transformado em variáveis de contagem, uma para cada categoria; e
- **candidatos.cand.sit.turno:** valor categórico da descrição da situação dos turnos das eleições da qual um servidor foi candidato. Como um mesmo servidor pode ser candidato em mais de um ano com situações diferentes nos diversos turnos, este atributo foi transformado em variáveis de contagem, uma para cada categoria.

3.3.2.4 Dimensão de Vínculos Societários

Conforme descrito no entendimento da Dimensão de Vínculos Societários na Seção 3.2.4, há informações de diversos contextos tratadas. Inicialmente, para os dados cadastrais de empresas com relação de vínculo com servidores, conforme descrito na Seção 3.2.4.1, os atributos levantados são exibidos a seguir, com *alias* definido em negrito:

- **vs.qtd.cnpjs:** quantidade de empresas diferentes com as quais um servidor possui ou possuiu vínculo;
- **vs.qtd.ae.sec.div:** quantidade de atividades econômicas da CNAE tidas como secundárias da empresa com a qual um servidor possui ou possuiu vínculo. Como um mesmo servidor pode possuir vínculo com mais de uma empresa, este atributo foi desdobrado em quatro, a saber: vs.qtd.ae.sec.div.total, vs.qtd.ae.sec.div.med, vs.qtd.ae.sec.div.max e vs.qtd.ae.sec.div.min;
- **vs.ae.princ.div:** valor categórico da atividade econômica da CNAE tida como principal da empresa com a qual um servidor possui ou possuiu vínculo. Como um mesmo servidor pode possuir vínculo com mais de uma empresa, pode estar

relacionado a mais de uma atividade principal, logo, este atributo foi transformado em variáveis de contagem, uma para cada categoria;

- **vs.nat.jur:** valor categórico do tipo de natureza jurídica da empresa com a qual um servidor possui ou possuiu vínculo. Como um mesmo servidor pode possuir vínculo com mais de uma empresa, pode estar relacionado a mais de uma natureza jurídica, logo, este atributo foi transformado em variáveis de contagem, uma para cada categoria;
- **vs.nat.jur.grupo:** valor categórico do grupo de naturezas jurídicas da empresa com a qual um servidor possui ou possuiu vínculo. Como um mesmo servidor pode possuir vínculo com mais de uma empresa, pode estar relacionado a mais de um grupo de natureza jurídica, logo, este atributo foi transformado em variáveis de contagem, uma para cada categoria;
- **vs.situacao:** valor categórico da situação da empresa com a qual um servidor possui ou possuiu vínculo. Como um mesmo servidor pode possuir vínculo com mais de uma empresa, pode estar relacionado com mais de uma situação, logo, este atributo foi transformado em variáveis de contagem, uma para cada categoria;
- **vs.porte.empresa:** valor categórico do porte da empresa com a qual um servidor possui ou possuiu vínculo. Como um mesmo servidor pode possuir vínculo com mais de uma empresa, pode estar relacionado a mais de um porte, logo, este atributo foi transformado em variáveis de contagem, uma para cada categoria;
- **vs.qtd.vinculos:** quantidade de vínculos vigentes ou não de um servidor com empresas diferentes;
- **vs.pc.participacao:** valor numérico indicador da porcentagem de participação de um servidor na empresa com a qual possui vínculo. Como um mesmo servidor pode possuir vínculo com mais de uma empresa, este atributo foi desdobrado em quatro, a saber: vs.pc.participacao.total, vs.pc.participacao.med, vs.pc.participacao.max e vs.pc.participacao.min;
- **vs.dias.vinculo:** tempo (em dias) em que um servidor esteve vinculado a uma empresa ou, para vínculos ainda ativos, tempo desde a entrada até uma data recente. Como um mesmo servidor pode estar ou já ter estado vinculado a mais de uma empresa, este atributo foi desdobrado em quatro, a saber: vs.dias.vinculo.total, vs.dias.vinculo.med, vs.dias.vinculo.max e vs.dias.vinculo.min;
- **vs.vigente.total:** quantidade de vínculos ativos de um servidor com empresas diferentes;

- **vs.qtd.como.responsavel:** quantidade de vínculos com designação de responsável para um servidor em empresas onde possui ou possui vínculo;
- **vs.ds.qualificacao:** valor categórico da descrição da qualificação de cada servidor na empresa com a qual possui ou possuiu vínculo. Como um mesmo servidor pode possuir mais de uma qualificação, este atributo foi transformado em variáveis de contagem, uma para cada categoria;
- **vs.qtd.func:** quantidade de funcionários da empresa com a qual um servidor possui ou possuiu vínculo. Como um mesmo servidor pode possuir vínculo com mais de uma empresa, este atributo foi desdobrado em quatro, a saber: vs.qtd.func.total, vs.qtd.func.med, vs.qtd.func.max e vs.qtd.func.min;
- **vs.sal.med.por.ano:** valor do salário médio por ano dos funcionários de empresa com a qual um servidor possui ou possuiu vínculo. Como um mesmo servidor pode possuir vínculo com mais de uma empresa, este atributo foi desdobrado em quatro, a saber: vs.sal.med.por.ano.total, vs.sal.med.por.ano.med, vs.sal.med.por.ano.max e vs.sal.med.por.ano.min;
- **vs.sal.min.ano:** valor do salário mínimo por ano dos funcionários de empresa com a qual um servidor possui ou possuiu vínculo. Como um mesmo servidor pode possuir vínculo com mais de uma empresa, este atributo foi desdobrado em quatro, a saber: vs.sal.min.por.ano.total, vs.sal.min.por.ano.med, vs.sal.min.por.ano.max e vs.sal.min.por.ano.min;
- **vs.sal.max.ano:** valor do salário máximo por ano dos funcionários de empresa com a qual um servidor possui ou possuiu vínculo. Como um mesmo servidor pode possuir vínculo com mais de uma empresa, este atributo foi desdobrado em quatro, a saber: vs.sal.max.por.ano.total, vs.sal.max.por.ano.med, vs.sal.max.por.ano.max e vs.sal.max.por.ano.min; e
- **vs.sal.total:** valor do salário total recebido em todos os anos dos funcionários de empresa com a qual um servidor possui ou possuiu vínculo. Como um mesmo servidor pode possuir vínculo com mais de uma empresa, este atributo foi desdobrado em quatro, a saber: vs.sal.total.total, vs.sal.total.med, vs.sal.total.max e vs.sal.total.min.

Para as informações de impedimentos de empresas, foram utilizadas as indicações da Seção 3.2.4.2 para criação dos atributos definidos a seguir, com *alias* explicitado em negrito:

- **vs.ceis.qtd.cnpjs:** quantidade de empresas diferentes cadastradas no CEIS com as quais um servidor possui ou possuiu vínculo;
- **vs.ceis.qtd.punicoes:** quantidade de punições diferentes de empresas cadastradas no CEIS com as quais um servidor possui ou possuiu vínculo;
- **vs.ceis.meses.sancao:** quantidade de meses onde cada sanção esteve vigente no CEIS para empresa vinculada a um servidor. Como um mesmo servidor pode estar vinculado a mais de uma empresa presente no CEIS, este atributo foi desdobrado em quatro, a saber: vs.ceis.meses.sancao.total, vs.ceis.meses.sancao.med, vs.ceis.meses.sancao.max e vs.ceis.meses.sancao.min;
- **vs.ceis.fund.legal:** valor categórico com fundamentação legal da sanção imputada a empresa com a qual um servidor possui ou possuiu vínculo. Como um mesmo servidor pode estar vinculado a mais de uma empresa presente no CEIS, pode estar relacionado a mais de uma fundamentação legal de sanção diferente, logo, este atributo foi transformado em variáveis de contagem, uma para cada categoria;
- **vs.ceis.tp.sancao:** valor categórico com descrição do tipo da sanção imputada a empresa com a qual um servidor possui ou possuiu vínculo. Como um mesmo servidor pode estar vinculado a mais de uma empresa presente no CEIS, pode estar relacionado a mais de um tipo de sanção diferente, logo, este atributo foi transformado em variáveis de contagem, uma para cada categoria;
- **vs.cepim.qtd.cnpjs:** quantidade de entidades diferentes cadastradas no CEPIM com as quais um servidor possui ou possuiu vínculo;
- **vs.cepim.qtd.punicoes:** quantidade de impedimentos, vigentes ou não, de entidades com as quais um servidor possui ou possuiu vínculo;
- **vs.cepim.qtd.vigentes:** quantidade de impedimentos vigentes de entidades com as quais um servidor possui ou possuiu vínculo; e
- **vs.cepim.dias.cepim:** quantidade de dias onde cada impedimento esteve vigente no CEPIM para entidade vinculada a um servidor. Como um mesmo servidor pode estar vinculado a mais de uma empresa presente no CEPIM, este atributo foi desdobrado em quatro, a saber: vs.cepim.dias.cepim.total, vs.cepim.dias.cepim.med, vs.cepim.dias.cepim.max e vs.cepim.dias.cepim.min.

Já em relação aos dados de doações eleitorais descritos na Seção 3.2.4.3, os seguintes atributos foram definidos, com *alias* explicitado em negrito:

- **vs.doacao.qtd.cnpjs:** quantidade de empresas diferentes, vinculadas a um servidor, que realizaram doações; e

- **vs.doacao.valor:** valor doado por empresas vinculadas a um servidor, conforme Seção 3.2.4.3. Como um mesmo servidor pode estar vinculado a mais de uma empresa que realiza ou realizou doações, este atributo foi desdobrado em quatro, a saber: vs.doacao.valor.total, vs.doacao.valor.med, vs.doacao.valor.max e vs.doacao.valor.min.

Para dados de recebimento de recursos públicos por parte de empresas, foram utilizadas as indicações da Seção 3.2.4.4 para tornar possível elencar os atributos apresentados a seguir, com *alias* explicitado em negrito:

- **vs.siconv.qtd.cnpjs:** quantidade de entidades ou empresas diferentes vinculadas a um servidor que participaram de convênios;
- **vs.siconv.qtd.convenio.total:** quantidade total de convênios realizados com participação de entidades ou empresas, como convenentes ou subcontratadas, vinculadas a um servidor;
- **vs.siconv.qtd.convenio.difer:** quantidade de convênios diferentes com participação de entidades ou empresas, como convenentes ou subcontratadas, vinculadas a um servidor;
- **vs.siconv.qtd.tipo.fornecedor:** quantidade de empresas diferentes, vinculadas a um servidor, que participaram de convênios como subcontratadas;
- **vs.siconv.qtd.tipo.participe:** quantidade de entidades diferentes, vinculadas a um servidor, que participaram de convênios como convenentes;
- **vs.ob.qtd.cnpjs:** quantidade de empresas diferentes, vinculadas a um servidor, que receberam OBC, OBB ou OBD;
- **vs.ob.qtd.obs:** quantidade de Ordens Bancárias – OBC, OBB ou OBD – recebidas por empresas vinculadas a um servidor; e
- **vs.ob.vl.ob:** valor de OB – OBC, OBB ou OBD – recebida por empresas vinculadas a um servidor. Como uma mesma empresa pode receber várias ordens bancárias e um mesmo servidor pode estar vinculado a mais de uma empresa, a partir do total recebido por cada empresa, este atributo foi desdobrado em quatro, a saber: vs.ob.vl.ob.total, vs.ob.vl.ob.med, vs.ob.vl.ob.max e vs.ob.vl.ob.min.

No que se refere à terceirização, foram utilizadas as indicações da Seção 3.2.4.5 e levantados os seguintes atributos, com *alias* explicitado em negrito:

- **vs.terceirizados.qtd.cnpjs:** quantidade de empresas diferentes, vinculadas a um servidor, que possuem funcionários empregados como terceirizados na Administração Pública Federal;
- **vs.terceirizados.qtd.ugs:** quantidade de unidades gestoras da Administração Pública Federal diferentes com empregados terceirizados de empresas vinculadas a um servidor;
- **vs.terceirizados.vl.mensal.salario:** valor do salário que terceirizados empregados de empresas vinculadas a um servidor recebem mensalmente. Como uma mesma empresa tem vários empregados com diversos salários e um mesmo servidor pode estar vinculado a mais de uma empresa, a partir do total de salários pagos mensalmente por cada empresa, este atributo foi desdobrado em quatro, a saber: vs.terceirizados.vl.mensal.salario.total, vs.terceirizados.vl.mensal.salario.med, vs.terceirizados.vl.mensal.salario.max e vs.terceirizados.vl.mensal.salario.min;
- **vs.terceirizados.vl.mensal.custo:** valor do salário mais encargos trabalhistas que terceirizados empregados de empresas vinculadas a um servidor recebem mensalmente. Como uma mesma empresa tem vários empregados com diversos custos e um mesmo servidor pode estar vinculado a mais de uma empresa, a partir do total de custos pagos mensalmente por cada empresa, este atributo foi desdobrado em quatro, a saber: vs.terceirizados.vl.mensal.custo.total, vs.terceirizados.vl.mensal.custo.med, vs.terceirizados.vl.mensal.custo.max e vs.terceirizados.vl.mensal.custo.min; e
- **vs.terceirizados.qtd.anos:** quantidade de anos diferentes havendo terceirizados empregados de empresas vinculadas a um servidor. Como um mesmo servidor pode estar vinculado a mais de uma empresa com terceirizados, pode estar relacionado a quantidades de anos diferentes, logo, este atributo foi desdobrado em quatro, a saber: vs.terceirizados.qtd.anos.total, vs.terceirizados.qtd.anos.med, vs.terceirizados.qtd.anos.max e vs.terceirizados.qtd.anos.min.

3.3.3 Análise de Variância e Correlação

Após a construção dos atributos – incluindo seus desdobramentos em atributos de valores totais, máximos, médios e mínimos, assim como atributos categóricos transformados em variáveis de contagem – tem-se apenas dados numéricos que representam contagens, sejam de quantidades ou de tempo, totalizando 968 atributos.

Contabilizando os dados por classes, como definido na Seção 3.3, caso hajam atributos que em uma mesma classe possuam frequência igual a zero, ou seja, tenham variância igual

a zero, não há necessidade de mantê-los, pois a maioria dos algoritmos de regressão não irão convergir e trarão estimativas de coeficientes iguais a zero ou infinito [36].

Dessa forma, foi calculada a variância por classe para todos os 968 atributos construídos e, após verificação daqueles com variância igual a zero em alguma classe, foram eliminados 412 atributos, dentre eles, pode-se citar diversas variáveis de contagem dos atributos siape.nivel.funcao, siape.situacao.funcional, vs.ds.qualificacao e vs.nat.jur.

É necessário observar ainda que atributos perfeitamente correlacionados podem ter sido adicionados de maneira acidental, ou as correlações podem ter surgido após as transformações realizadas na limpeza e construção de dados. Como atributos com alta correlação não devem ser mantidos nas análises de regressão, conforme discutido na Seção 2.6.3, calculou-se a correlação de todos os 556 atributos mantidos após análise de variância.

Entre os atributos analisados, apenas quatro – dois pares – retornaram perfeita correlação, são eles: vs.ceis.fund.legal.Art7Lei105202002 com vs.ceis.tp.sancao.ImpedimentoLeidoPregao, e candidatos.sit.candidatura.1NE com candidatos.nr.ano.eleicao.1994. Foram eliminados, por escolha dos especialistas da DIE, os primeiros relacionadas em cada par. Dessa forma, restaram 554 atributos dos 968 iniciais após a análise de variância e correlação.

3.3.4 Separação de Dados

Com os atributos construídos devidamente filtrados após análise de variância e correlação, tem como conjunto de dados registros de 5.170 corruptos e 285.620 não corruptos – dados claramente desbalanceados, conforme definido na Seção 2.2. Como já apresentado na Seção 3.3, os corruptos são aqueles servidores públicos federais presentes no CEAF. Já para os não corruptos, a quantidade retornada decorre do limite superior de 300.000 dado à consulta realizada nas bases de dados seguido da eliminação de corruptos do resultado e limpeza de dados.

Com o conjunto citado é realizada a separação inicial em dados de treino (DT) – para seleção de atributos e modelagem – e teste (DTE) – de forma a possibilitar uma verificação final, após todos os procedimentos de modelagem e avaliação, em dados que não foram utilizados em nenhuma atividade, servindo portanto como dados reais mas com valores definidos de classe.

Com essa finalidade, para construir o DTE separou-se de maneira aleatória 10%[20] dos corruptos e a mesma quantidade de não corruptos através de *under-sampling* de forma a manter os dados balanceados, como apresentado na Seção 2.2. Portanto, como dados de treino (DT) foram selecionados os 90% de corruptos e os dados de não corruptos

[20]Valor percentual de 10% para DTE foi escolhido arbitrariamente. Assim, pode ser útil variar a porcentagem para comparação de resultados em trabalhos futuros.

restantes. Dessa forma, como pode ser visualizado na Figura 3.1, os dados de treino (DT) serão aqueles considerados durante todas fases seguintes, até o momento onde, após a modelagem e avaliação de um modelo dito final, os dados de teste (DTE) serão verificados para confirmação de todo o processo de mineração de dados, ratificando-o, ou sua refutação, que exige um reinício desde a fase inicial.

3.3.5 Ajustes de Atributos para Seleção

A fase de preparação de dados, como realizada neste trabalho, contempla inclusive o ajuste de todos os atributos construídos para que em momento posterior aqueles mais relevantes em termos de corruptibilidade sejam selecionados em processo que inclui atividades de modelagem, a ser explicitado na Seção 3.4.1.

Com esse intuito, como visto na Figura 3.1, inicialmente o conjunto DT, separado na Seção 3.3.4, é dividido em treino (DTS) e validação (DVS) para o processo de seleção de atributos. Para tal, foram amostrados aleatoriamente 80%[21] dos corruptos de DT para DTS, juntamente com uma mesma quantidade de não corruptos através de *under-sampling*, de forma a manter os dados balanceados, como apresentado na Seção 2.2. Consequentemente, para DVS, tem-se 20% dos corruptos de DT e o mesmo quantitativo de não corruptos.

Em seguida, nas próximas seções, serão descritas as atividades realizadas para ajustar os dados analisando interações entre os atributos e um conjunto pré-determinado deles, incluindo ainda a verificação de não linearidade. Finalmente, após tal verificação, os atributos observados como possuindo comportamento quadrático são elevados ao quadrado para uso na modelagem e os atributos restantes são discretizados fazendo uso de três algoritmos diferentes.

3.3.5.1 Interações entre Atributos

Um fenômeno comumente encontrado no mundo é que a influência da ação simultânea de dois atributos não é a mesma que a soma das influências independentes dos dois atributos [69]. Assim, em modelos aditivos, como os MLG apresentados na Seção 2.4.2, é necessário identificar interações entre atributos, usualmente introduzidos pelo produto dos mesmos [69].

Buscando investigar a interação entre diferentes atributos, por ser inviável devido a limitações de tempo e computação a verificação de todas as combinações possíveis de atributos, os especialistas da DIE selecionaram entre os 554 atributos disponíveis alguns que seriam possivelmente mais representativos em interações com os demais.

[21] Valor percentual de 80% para DTS foi escolhido arbitrariamente. Assim, pode ser útil variar a porcentagem para comparação de resultados em trabalhos futuros.

Por conseguinte, foram identificados 36 atributos de forma a serem criadas novos atributos a partir da interação dos selecionados com todos os outros atributos. Separando-se os atributos que criarão interações pelas dimensões definidas na Seção 3.2, temos:

- **Dimensão Funcional**
  - siape.sal.bruto.max
  - siape.dias.cargo.total
  - siape.situacao.funcional.ATIVOPERMANENTE
  - siape.situacao.funcional.APOSENTADO
  - siape.nivel.funcao.DAS1011
  - siape.nivel.funcao.DAS1012
  - siape.nivel.funcao.DAS1013
  - siape.nivel.funcao.DAS1014
  - siape.nivel.funcao.DAS1015
  - siape.nivel.funcao.DAS1016
  - siape.nivel.funcao.FCI0001
  - tcuirregulares.qtd.total.contas.julg.irreg
  - ativa.qtd.const
  - ativa.qtd.const.por.os
  - ativa.tipo.const.FALHAMEDIA
  - ativa.tipo.const.FALHAGRAVE
  - cgie.tipo.envolvimento.Investigado
- **Dimensão Política**
  - filiados.dias.filiacao.total
  - filiados.qtd.total.filiacoes
- **Dimensão de Vínculos Societários**
  - vs.qtd.cnpjs
  - vs.ae.princ.div.ATIVIDADESDESEDESDEEMPRESASEDECONSULTORIA-EMGESTAOEMPRESARIAL
  - vs.ae.princ.div.ATIVIDADESDEORGANIZACOESASSOCIATIVAS
  - vs.ae.princ.div.CONSTRUCAODEEDIFICIOS

- vs.ae.princ.div.OBRASDEINFRAESTRUTURA
- vs.ae.princ.div.SERVICOSESPECIALIZADOSPARACONSTRUCAO
- vs.nat.jur.grupo.ENTIDADESSEMFINSLUCRATIVOS
- vs.qtd.vinculos
- vs.vigente.total
- vs.qtd.como.responsavel
- vs.ds.qualificacao.CONTADOR
- vs.ceis.qtd.punicoes
- vs.cepim.qtd.punicoes
- vs.doacao.valor
- vs.siconv.qtd.convenio.total
- vs.ob.vl.ob.total
- vs.terceirizados.qtd.ugs

Assim, após as interações dos atributos selecionados com os outros, tem-se um conjunto de dados com os 554 atributos originais, os 36 selecionados combinados 2 a 2, gerando mais 630 atributos, além das interações dos 36 com todos os outros 518 não selecionados, que gera 18648 atributos. Portanto, tem-se um total de 19832 atributos após as interações.

3.3.5.2 Análise de Variância, Correlação e Não Linearidade

De modo similar ao realizado na Seção 3.3.3, foi analisada a variância de todos os atributos obtidos após realizadas as interações, eliminando 8056 atributos com variância igual a zero em alguma das classes. A análise de correlação também foi realizada mas com uma abordagem um pouco diferente daquela feita na Seção 3.3.3.

Inicialmente foi criada a matriz de correlação dos 11776 atributos restantes após análise de variância, adicionado ainda a correlação deles com a coluna da classe indicando a corruptibilidade – 0 para não corruptos e 1 para corruptos. Com os valores gerados, foram filtrados os pares de atributos com correlação em módulo maior ou igual a 0.70 – valor considerado com regra geral de alta correlação [65]. Em seguida, a matriz resultante da filtragem foi ordenada de maneira decrescente em relação à correlação dos atributos presentes com a classe.

A partir do resultado da matriz filtrada e ordenada, as linhas foram percorridas a partir dos maiores valores de correlação entre cada atributo e a classe. Em cada linha, o atributo com maior correlação com a classe foi mantido e os restantes eliminados. Com esse algoritmo foram eliminados 8845 atributos que possuíam correlação em módulo maior

ou igual a 0.70, restando, portanto, 2931 atributos. Tal abordagem foi construída pela necessidade de remoção de atributos correlacionados para buscar evitar o problema de colinearidade, como explicitado na Seção 2.6.3, principalmente devido ao fato de não ser possível eliminar colinearidade analisando todos as combinações possíveis de grupos de atributos, devido à grande quantidade envolvida neste trabalho. Assim, a heurística de correlação de cada atributo com sua classe – apesar de não ser totalmente refletida na regressão devido às interações entre os atributos de um modelo – serve como técnica para tentar manter os atributos teoricamente mais significativos – considerando a correlação com a classe – observando apenas a influência isolada de cada atributo[22].

É importante mencionar que há outras técnicas úteis para resolver a questão de atributos correlacionados, como PCA (*Principal Components Analysis* [67]). Tal técnica consegue reduzir o número de atributos construindo variáveis derivadas criadas a partir de uma combinação dos atributos existentes. Apesar de bastante usada, optou-se por não utilizar PCA neste trabalho devido ao fato da criação de atributos combinados diminuir a interpretabilidade do modelo resultante, dificultando a análise conjunta das regras geradas com os especialistas em combate à corrupção.

Como visto na Seção 2.4.2, ao adicionar atributos em um algoritmo de regressão aditivo, é importante verificar se o comportamento assumido se reflete em cada atributo. Assim, por exemplo, caso um atributo se comporte de maneira quadrática, é necessário que o mesmo seja introduzido como potência de 2 para a modelagem. Além disso, em modelos aditivos, são desejáveis atributos de comportamento monótono – ou seja, sempre crescente ou decrescente – de maneira que um atributo de comportamento não-monótono requer transformações para sua melhor adequação, como, por exemplo, discretização [69].

Assim, utilizando o método de regressão local LOESS, citado na Seção 2.4.1 no conjunto DTS, com os resultados aplicados em uma função logit, para tentar refletir uma resposta logística[23], foi gerado um gráfico com os valores de logit para os valores dos atributos para cada um dos 2931, de modo a analisar seus comportamentos.

Os gráficos gerados foram analisados individualmente observando visualmente as curvas geradas e optou-se por checar, devido a restrições de tempo, apenas comportamentos quadráticos. Após a análise gráfica individual dos atributos, 28 foram selecionados para serem elevados ao quadrado por terem comportamento aproximadamente quadrático, como, por exemplo, o atributo vs.qtd.ae.sec.div.total, com gráfico mostrado na Figura 3.4.

Os 2903 atributos não selecionados para serem elevados ao quadrado serão discretizados fazendo uso de três algoritmos diferentes, como mostrado na próxima seção. Apesar

[22]Heurística foi escolhida para este trabalho sem ter sido analisada em outros trabalhos. Assim, pode ser útil tratar o problema de atributos correlacionados com outras técnicas em trabalhos futuros.

[23]Checking functional form in logistic regression using loess plots: http://thestatsgeek.com/2014/09/13/checking-functional-form-in-logistic-regression-using-loess/

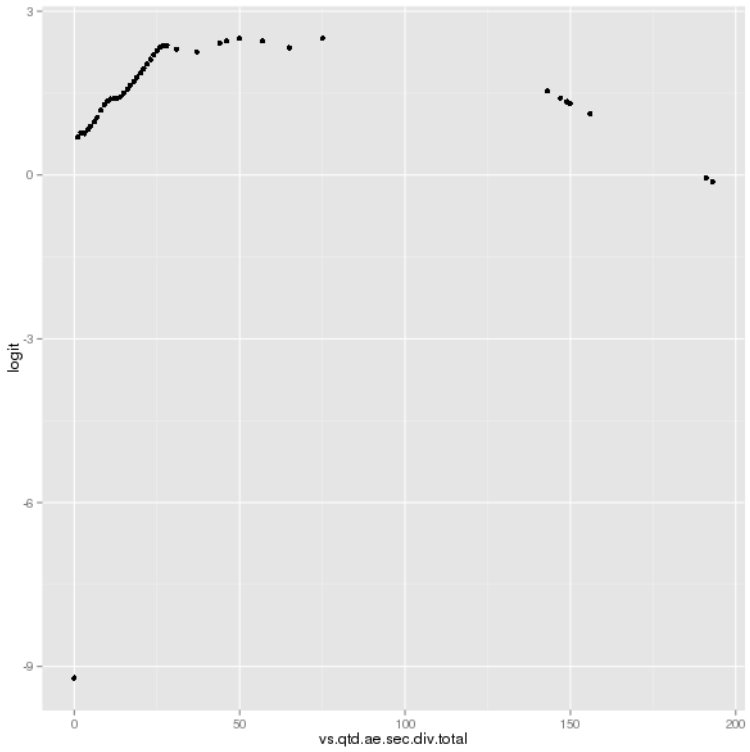

Figura 3.4: Comportamento do atributo vs.qtd.ae.sec.div.total

de alguns deles terem comportamento linear, optou-se pela discretização para o modelo ter como resultado variáveis binárias – ou seja, com valores 1 para existência de um atributo em uma dada faixa definida na discretização e valores 0 para inexistência – facilmente explicáveis em um contexto de apresentação de conhecimento descoberto[24].

3.3.5.3 Discretização

Seguindo os algoritmos de discretização CAIM, MDLP e ModChi2, mencionados na Seção 2.5, os atributos não elevados ao quadrado foram transformados de variáveis contínuas em binárias para faixas de valores escolhidas pelos algoritmos. Em outras palavras, um dado atributo que, por exemplo, seja separado em duas faixas de valores, será desdobrado em dois atributos binários, com valor 1 para existência do valor em uma das faixas e valor 0 para inexistência.

Dessa forma, foram gerados três conjuntos de dados diferentes para o conjunto DTS, um conjunto para cada discretização usada. O conjunto de dados com os atributos discretizados pelo algoritmo CAIM retornou 6399 atributos binários. Já o conjunto discretizado com MDLP retornou 877 atributos binários e o que utilizou ModChi2 restou com 880. A

[24]Não realizar discretização de atributos lineares para algoritmos aditivos pode ser útil, sendo assim válido considerar em trabalhos futuros.

questão de haver conjuntos de dados com um quantitativo de atributos menor do que os 2903 originais é explicada pelo fato das funções criadas para discretização eliminaram os atributos constantes gerados, ou seja, aqueles atributos com o mesmo valor para todos os registros – por exemplo, um atributo que corresponde a uma faixa onde não há registros em nenhum dado, possuindo 0 como valor em todos os registros.

Após a criação dos três conjuntos de dados diferentes, cada um deles passou pela mesma análise de variância por classe e correlação – com limiar de 0.70 – relatada na Seção 3.3.5.2. Depois de tal atividade, os conjuntos para CAIM, MDLP e ModChi2 acabaram com, respectivamente, 2839, 335 e 413 atributos binários – incluindo na contagem os 28 atributos que foram elevados ao quadrado.

3.4 Modelagem

A fase de modelagem consiste na construção de modelos a partir dos conjuntos de dados utilizando métodos de regressão definidos. Pelo fato de, neste trabalho, a seleção de atributos incluir a criação de modelos para definição dos atributos mais adequados, a seção a seguir descreve os passos realizados em tal atividade. Feita a escolha dos atributos, os dados definidos serão utilizados para a construção de modelos que, após avaliação, resultarão nas regras de avaliação de risco de corrupção.

3.4.1 Seleção de Atributos

Para a realização da seleção de atributos, as atividades descritas a seguir serão realizadas para cada um dos três conjuntos de dados gerados na Seção 3.3.5.3, de modo que, ao término, seja possível comparar os resultados e selecionar o grupo de atributos mais adequado.

Inicialmente, cada conjunto de dados passa por uma regressão com regularização, mais especificamente utilizando *Adaptive Lasso*, conforme explicitado na Seção 2.4.3. Assim, primeiro executa-se no conjunto de dados DTS a Regressão *Ridge* com *10-fold cross-validation* através do R usando o pacote glmnet[25], para extrair os coeficientes estimados. Tais estimativas são usadas para construir o vetor adaptativo de pesos $\hat{\omega}_j$, definido na Equação 2.10, de modo que o vetor de coeficientes resultante da Regressão *Ridge* é o $\hat{\beta}_j^{ini}$ da citada equação – considerando-se a constante γ da mesma equação igual a 1[26].

Dessa forma, o vetor adaptativo de pesos é o fator de penalidade introduzido na execução do *Adaptive Lasso* com *10-fold cross-validation* através do R, com parâmetro alpha igual a 1. É válido frisar que, devido ao modo de construção da penalização do

[25] Pacote glmnet: https://cran.r-project.org/web/packages/glmnet/

[26] Valor de γ escolhido arbitrariamente, assim pode ser útil analisar outros valores em trabalhos futuros.

Adaptive Lasso, os coeficientes podem ter coeficientes estimados iguais a zero, reduzindo portanto o número de atributos dos modelos construídos.

Assim, com os atributos selecionados através do *Adaptive Lasso*, é executado o método de regressão logística com *10-fold cross-validation* através do R com o pacote caret[27] [44]. Portanto, após essas execuções cada um dos três conjunto de dados gerados terá um modelo de regressão logística.

Com o intuito de escolher o grupo de atributos mais adequado para prosseguir para a fase seguinte na mineração de dados, os resultados do *10-fold cross-validation* com o método de regressão logística são comparados para os três conjuntos de dados, gerando a Tabela 3.1 a seguir.

Tabela 3.1: Resultados dos modelos gerados na seleção de atributos

Discretização	Nº de atributos	**Sensibilidade**	**Precisão**	**Acurácia**	**F-Measure**
CAIM	316	0.9076	0.8447	0.8704	0.8750
MDLP	120	0.8600	0.8426	0.8497	0.8512
ModChi2	100	0.8049	0.7968	0.7998	0.8008

Apesar dos resultados observados para o conjunto de dados com discretização via CAIM terem sido melhores, é recomendável tentar minimizar a quantidade de atributos considerados na modelagem [36], pois o modelo resultante tende a ser numericamente estável e pode ser adotado mais facilmente, além do fato de que quanto mais atributos em um modelo, maiores os erros estimados e mais dependente fica o modelo em relação aos dados [36].

Assim, optou-se por selecionar os atributos do conjunto de dados discretizado com MDLP, por terem obtido resultados próximos aos discretizados com CAIM mas mantendo quase três vezes menos atributos.

De forma a ratificar a escolha do modelo gerado com os dados discretizados com MDLP, o mesmo foi executado no conjunto de validação DVS separado na Seção 3.3.5, gerando os resultados da Tabela 3.2 a seguir, onde também foram adicionados os resultados obtidos anteriormente no conjunto de treino DTS para fins de comparação.

Tabela 3.2: Resultados do modelo para dados discretizados com MDLP

Dados	**Sensibilidade**	**Precisão**	**Acurácia**	**F-Measure**
DTS	0.8600	0.8426	0.8497	0.8512
DVS	0.8131	0.7977	0.8034	0.8053

[27]Pacote caret: https://cran.r-project.org/web/packages/caret/

Portanto, observa-se que os resultados de validação no conjunto DVS foram próximos dos de treino (DTS), confirmando a validade do modelo e corroborando a escolha dos atributos presentes no mesmo.

3.4.2 Construção de Modelos

Inicialmente para a construção de modelos e posterior avaliação dos resultados, os dados de treino (DT) selecionados na Seção 3.3.4 foram ajustados de modo a deixá-los na mesma estrutura do conjunto de dados discretizado com MDLP, conforme definido na Seção 3.4.1.

Em seguida, foi utilizado processo de reamostragem com *Bootstrap* [68], conforme visto na Seção 2.3.1. Tal processo busca atacar inicialmente a questão de haver servidores de corruptibilidade desconhecida no conjunto de não corruptos, como apontado na Seção 3.3, a partir da criação de várias subamostras dos dados com a geração de um modelo diferente para cada subamostra, com o intuito de minimizar a influência geral de servidores possivelmente corruptos no grupo de não corruptos. Com o uso de métodos de regularização, busca-se ainda obter um subconjunto dos atributos. Assim, tem-se o *Bootstrap* de acordo os passos a seguir:

1. Seleciona-se uma amostra aleatória sem reposição de 50% dos corruptos de DT, denominada amostra **A1**;
2. Cria-se duas subamostras a partir de **A1**:
   (a) **Treino (DTA)**: com uma seleção aleatória de 50% dos corruptos de **A1** e a mesma quantidade de não corruptos selecionada aleatoriamente;
   (b) **Validação (DVA)**: 50% restantes dos corruptos de **A1** e outra amostra aleatória de não corruptos de mesma quantidade.
3. Utilizando a subamostra DTA, o método de regressão *Adaptive Lasso* é aplicado – seguindo exatamente os mesmos passos mostrados na Seção 3.4.1 – usando os coeficientes estimados na Regressão *Ridge* no vetor adaptativo de pesos;
4. Ainda com a subamostra DTA, mas usando apenas os atributos selecionados pelo *Adaptive Lasso*, cria-se um modelo executando o método de regressão logística, igualmente ao mencionado na Seção 3.4.1;
5. Finalmente, o modelo criado é executado nos dados DVA de forma a gerar resultados de métricas de validação.

Em uma primeira execução o processo de modelagem definido, buscando sempre amostras aleatórias **A1** dos dados, foi realizado em 5.000 iterações – esse número foi definido

arbitrariamente, sendo sua suficiência atestada nos resultados das distribuições das estimativas obtidas no processo, onde distribuições normais ratificam uma quantidade suficiente de iterações [68].

Buscando reduzir ainda mais a quantidade de atributos – com o intuito de simplificar o modelo final, melhorar a capacidade de representação do conhecimento e diminuir a dependência do modelo em relação aos dados [36] – assim como tentando favorecer a convergência das estimativas dos coeficientes e das métricas de validação, foi analisada a frequência dos atributos nos modelos gerados em cada iteração. Assim, dos 120 atributos inicialmente analisados, 32 deles estavam presentes em mais do que 50% dos modelos em iterações diferentes.

Com os 32 atributos fixados, o processo de modelagem foi ligeiramente modificado, apenas removendo o passo 3 de aplicação do *Adaptive Lasso* e usando no passo 4 os próprios atributos fixados. Após esta adaptação, o processo foi executado novamente em 1.000 iterações – quantidade definida seguindo a mesma lógica do número de iterações escolhido para a primeira execução do processo.

Como mencionado na Seção 2.3.1, para verificar a estabilização dos coeficientes estimados, é necessário observar se a distribuição dos valores obtidos nas iterações formam aproximadamente uma distribuição Normal. Dessa forma, para cada coeficiente estimado foi gerado um histograma e todas as distribuições foram observadas individualmente de forma a ratificar a visualização da distribuição Normal. É possível observar, por exemplo, o histograma para os valores estimados do coeficiente siape.situacao.funcional.ATIVO-PERMANENTE.]-99999999999, 0.5] na Figura 3.5 a seguir.

Calculando-se a média das estimativas nas iterações tem-se os valores finais das estimativas dos 32 coeficientes e da constante *Intercept* juntamente com os seus erros padrão – tais resultados estão delineados na Tabela 3.3. O modelo final é, portanto, construído usando os coeficientes finais obtidos via *Bootstrap*.

Tabela 3.3: Estimativas do modelo final

Atributo	Coeficiente	Erro Padrão
ativa.qtd.const.por.os..X..siape.situacao.funcional.NA.]-99999999999, 0.5]	-6.1397	0.2259
ativa.qtd.os.]-99999999999, 0.5]	-1.5330	0.0309
cgie.tipo.envolvimento.Investigado.]-99999999999, 0.5]	-2.4511	0.0484
filiados.qtd.total.filiacoes.]-99999999999, 0.5]	-1.0211	0.0048
pad.tp.penalidade.Advertencia.]-99999999999, 0.5]	-10.8765	0.2421
pad.tp.penalidade.Suspensao.]-99999999999, 0.5]	-10.3271	0.2448
siape.dias.cargo.min.]10254.5, 99999999999]	-14.4764	0.0305
siape.dias.cargo.min.]3721, 3769]	-3.8366	0.1482
siape.dias.cargo.min.]4675.5, 5460]	-1.2089	0.0124
siape.dias.cargo.min.]714.5, 739.5]	0.5480	0.0052
siape.dias.cargo.min.]739.5, 2797]	-3.6902	0.1886
siape.dias.cargo.total..X..siape.situacao.funcional.CELETISTAEMPREGADO.]7.5, 278]	-5.9476	0.2332
siape.dias.cargo.total..X..siape.qtd.atividades.]9944, 99999999999]	-1.0460	0.0081
siape.dias.cargo.total..X..siape.qtd.orgaos.funcao.]742.5, 2792.5]	0.2227	0.0055
siape.dias.cargo.total..X..siape.qtd.orgaos.funcao.]3251.5, 3448]	-1.3474	0.0122
siape.situacao.funcional.APOSENTADO..X..siape.dias.cargo.min.]2881.5, 3905.5]	-1.1180	0.0160
siape.situacao.funcional.APOSENTADO..X..siape.qtd.cargos.]3.5, 4.5]	-1.1666	0.0466
siape.situacao.funcional.APOSENTADO..X..tcuirregulares.qtd.total.contas.julg.irreg.]-99999999999, 0.5]	-4.7617	0.3217
siape.situacao.funcional.ATIVOPERMANENTE.]-99999999999, 0.5]	-1.6333	0.0051
siape.situacao.funcional.INSTITUIDORPENSAO.]-99999999999, 0.5]	2.5531	0.0417
siape.situacao.funcional.NA.]0.5, 1.5]	-0.7017	0.0089
siape.nivel.funcao.DAS1011.]-99999999999, 0.5]	-1.3153	0.0159
siape.nivel.funcao.FCI0001.]-99999999999, 0.5]	-5.6258	0.2203
siape.nivel.funcao.FGR0002.]-99999999999, 0.5]	-1.2885	0.0165
siape.qtd.cargos.]-99999999999, 0.5]	2.4852	0.0103
siape.qtd.orgaos.cargo.]0.5, 1.5]	-0.5695	0.0048
siape.sal.bruto.max..X..siape.situacao.funcional.CEDIDOSUSLEI8270.]-99999999999, 8828.24]	1.7695	0.0589
siape.sal.bruto.max..X..siape.nivel.funcao.NA.]2972.915, 3011.965]	-14.7420	0.0189
siape.sal.bruto.max.]-99999999999, 11.92]	1.3252	0.0048
siape.sal.bruto.max.]11.92, 334.435]	1.8205	0.0161
tcuirregulares.qtd.total.contas.julg.irreg.]0.5, 99999999999]	6.8981	0.2056
vs.pc.participacao.med.]34.8333333333333, 99999999999]	0.6744	0.0062
Intercept	41.0775	0.5933

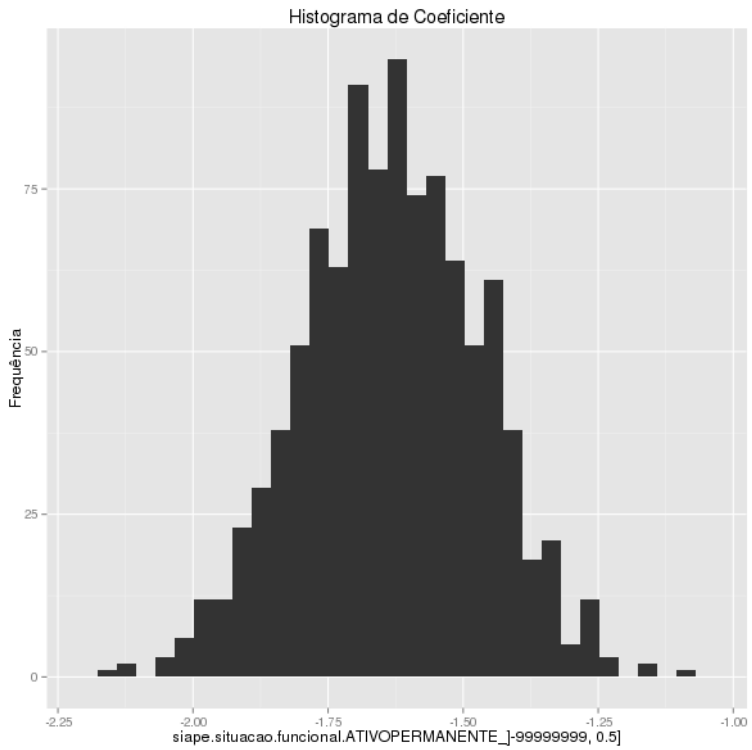

Figura 3.5: Histograma do Coeficiente siape.situacao.funcional.ATIVOPERMANENTE.]-99999999999, 0.5]

Capítulo 4

Resultados

Neste Capítulo serão apresentados os resultados obtidos após a efetivação da solução proposta seguindo as fases apontadas no modelo de referência CRISP-DM. A seguir serão delineadas as fases de Avaliação e Implantação. Na primeira, atividades de validação do modelo final obtido na Seção 3.4.2 são realizadas, buscando atestar a adequação do conhecimento gerado a partir dos dados, aplicando testes estatísticos e obtendo métricas de validação. Além disso, as regras geradas são analisados e alguns casos pontuais são estudados.

Finalmente na Implantação, serão indicadas tarefas para colocar em prática o conhecimento adquirido durante este projeto de mineração de dados, assim como meios de apresentação dos resultados para os interessados.

4.1 Avaliação

Na fase de Avaliação, utilizando como base o modelo final de regressão logística criado na Seção 3.4.2, com os coeficientes indicados na Tabela 3.3, algumas atividades de validação são executadas com o intuito de verificar a qualidade do modelo. As métricas de validação da execução do modelo final são apresentadas, assim como são realizados alguns procedimentos estatísticos como o Teste de Wald e o Teste de Hosmer-Lemeshow. Além disso, o próprio conhecimento obtido é analisado buscando ratificar sua adequação, a partir do estudo das regras geradas e da aplicação destas em alguns casos individuais.

4.1.1 Métricas de Validação

Na etapa de Construção de Modelos, delineada na Seção 3.4.2, o modelo final foi obtido com 32 atributos e uma constante. Como discutido, para tal, em um último passo, foram

realizadas 1.000 iterações de modelagem. Em cada iteração, as métricas de validação – precisão, sensibilidade, especificidade e acurácia – foram contabilizadas.

Assim, como mencionado na Seção 2.3.1, para verificar a estabilização dos resultados das métricas de validação, é necessário observar se a distribuição dos valores obtidos nas iterações realizadas na Seção 3.4.2 formam aproximadamente uma distribuição Normal. Dessa forma, é possível observar na Figura 4.1 a seguir os histogramas de todas as quatro métricas de validação obtidas durante a modelagem.

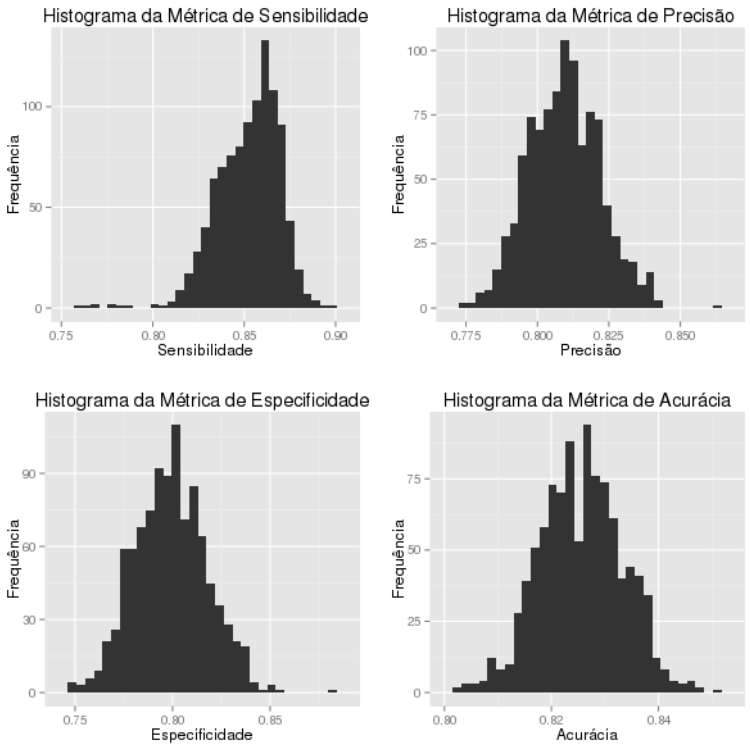

Figura 4.1: Histogramas das métricas de validação

Observa-se nos histogramas que todas as métricas possuem distribuição aproximadamente Normal, o que ratifica a validade do processo realizado na modelagem, assim como o modelo final gerado.

Dessa forma, a partir dos resultados obtidos via *Bootstrap*, como mostrado na Seção 3.4.2, têm-se os valores finais das métricas de validação calculando-se a média dos resultados nas iterações – tais resultados, juntamente com os seus erros padrão, estão delineados na Tabela 4.1.

É possível notar, portanto, que os resultados das métricas de validação do modelo final obtidos nas iterações de modelagem foram satisfatórios.

Tabela 4.1: Métricas de validação do modelo final

Métrica	Estimativa	Erro Padrão
Sensibilidade	0.8521	0.0005
Especificidade	0.7989	0.0005
Precisão	0.8094	0.0003
Acurácia	0.8255	0.0002

4.1.2 Teste de Wald

Como explicitado na Seção 2.6.1, para verificar se os coeficientes encontrados no modelo final para cada atributo são estatisticamente significativos, é realizado o teste de Wald [32]. As estimativas dos coeficientes, assim como de seus respectivos erros padrão, foram mostrados na Tabela 3.3. Assim, com os testes de hipótese realizados a partir da aplicação de tais estimativas na Equação 2.12 – considerando hipótese nula como cada coeficiente individualmente igual a zero – foram obtidos os resultados de p-valor mostrados na Tabela 4.2.

Como todos os resultados de p-valor foram menores do que 0.05, pode-se rejeitar a hipótese nula de coeficiente igual a zero para todos os atributos obtidos com nível de confiança de 95%, ratificando a presença dos mesmos no modelo gerado.

4.1.3 Teste de Hosmer-Lemeshow

Como medida de qualidade de ajuste do modelo, com a finalidade de avaliar o modelo final gerado a partir da comparação entre os resultados estimados e observados, o teste de Hosmer-Lemeshow [35] foi utilizado.

Foi separado um conjunto de dados especificamente para esse teste, utilizando uma amostra aleatória do conjunto de dados DT separados na Seção 3.3.4 com 80% dos corruptos e a mesma quantidade de não corruptos, de modo a manter o conjunto balanceado, também obtida aleatoriamente.

Assim, o modelo final foi executado nos dados separados para este teste, de forma a obter os resultados estimados de corruptibilidade, a serem comparados com os originalmente classificados.

Utilizando o parâmetro de $g = 5$[1] grupos no Teste de Hosmer-Lemeshow, obteve-se:

$$P - valor = 0.08244 \tag{4.1}$$

[1] O número de grupos para o Teste de Hosmer-Lemeshow foi escolhido arbitrariamente, assim pode ser útil testar outros valores em trabalhos futuros.

Tabela 4.2: Teste de Wald para coeficientes do modelo final

Atributo	P-valor
ativa.qtd.const.por.os..X..siape.situacao.funcional.NA.]-99999999999, 0.5]	1.258347e-162
ativa.qtd.os.]-99999999999, 0.5]	0
cgie.tipo.envolvimento.Investigado.]-99999999999, 0.5]	0
filiados.qtd.total.filiacoes.]-99999999999, 0.5]	0
pad.tp.penalidade.Advertencia.]-99999999999, 0.5]	0
pad.tp.penalidade.Suspensao.]-99999999999, 0.5]	0
siape.dias.cargo.min.]10254.5, 99999999999]	0
siape.dias.cargo.min.]3721, 3769]	9.280723e-148
siape.dias.cargo.min.]4675.5, 5460]	0
siape.dias.cargo.min.]714.5, 739.5]	0
siape.dias.cargo.min.]739.5, 2797]	3.281178e-85
siape.dias.cargo.total..X..siape.situacao.funcional.CELETISTAEMPREGADO.]7.5, 278]	2.233924e-143
siape.dias.cargo.total..X..siape.qtd.atividades.]9944, 99999999999]	0
siape.dias.cargo.total..X..siape.qtd.orgaos.funcao.]742.5, 2792.5]	0
siape.dias.cargo.total..X..siape.qtd.orgaos.funcao.]3251.5, 3448]	0
siape.situacao.funcional.APOSENTADO..X..siape.dias.cargo.min.]2881.5, 3905.5]	0
siape.situacao.funcional.APOSENTADO..X..siape.qtd.cargos.]3.5, 4.5]	3.038624e-138
siape.situacao.funcional.APOSENTADO..X..tcuirregulares.qtd.total.contas.julg.irreg.]-99999999999, 0.5]	1.445767e-49
siape.situacao.funcional.ATIVOPERMANENTE.]-99999999999, 0.5]	0
siape.situacao.funcional.INSTITUIDORPENSAO.]-99999999999, 0.5]	0
siape.situacao.funcional.NA.]0.5, 1.5]	0
siape.nivel.funcao.DAS1011.]-99999999999, 0.5]	0
siape.nivel.funcao.FCI0001.]-99999999999, 0.5]	7.889012e-144
siape.nivel.funcao.FGR0002.]-99999999999, 0.5]	0
siape.qtd.cargos.]-99999999999, 0.5]	0
siape.qtd.orgaos.cargo.]0.5, 1.5]	0
siape.sal.bruto.max..X..siape.situacao.funcional.CEDIDOSUSLEI8270.]-99999999999, 8828.24]	5.040985e-198
siape.sal.bruto.max..X..siape.nivel.funcao.NA.]2972.915, 3011.965]	0
siape.sal.bruto.max.]-99999999999, 11.92]	0
siape.sal.bruto.max.]11.92, 334.435]	0
tcuirregulares.qtd.total.contas.julg.irreg.]0.5, 99999999999]	1.097015e-246
vs.pc.participacao.med.]34.8333333333333, 99999999999]	0
Intercept	0

Como visto na Seção 2.6.2, a hipótese nula é que o modelo é adequado. Logo, como o resultado de p-valor foi maior que 0.05, a hipótese nula de adequação do modelo final não é rejeitada, com nível de confiança de 95%, o que corrobora sua validade na previsão de corruptibilidade.

4.1.4 Resultados para Dados de Teste

Como apresentado na Seção 3.3.4, o conjunto de dados de teste (DTE) separado inicialmente é usado para confirmação de todo o processo de mineração de dados, ratificando-o, ou para sua refutação, que exige um reinício desde a fase inicial. Como tais dados foram separados antes dos ajustes dos atributos, fez-se necessário seguir o mesmo processo para tais dados, discretizando-os com o algoritmo MDLP e selecionando apenas os atributos utilizados na modelagem.

Dessa forma, o modelo final obtido na Seção 3.4.2 foi executado nos dados de teste ajustados, tornando possível obter os valores de métricas para esses dados. Tais resultados são exibidos na Tabela 4.3, juntamente com as métricas apresentadas na Seção 4.1.1 para fins de comparação.

Tabela 4.3: Métricas do modelo final

Métrica	Treino	**Teste**
Sensibilidade	0.8521	**0.8607**
Especificidade	0.7989	**0.8066**
Precisão	0.8094	**0.8165**
Acurácia	0.8255	**0.8337**

Como pode-se observar, os resultados para os dados de teste foram tão satisfatórios quanto aqueles obtidos para o conjunto de treino, o que confirma a validade do modelo final gerado.

4.1.5 Análise de Regras Geradas

A partir do modelo final gerado na Seção 3.4.2, analisando o significado dos atributos escolhidos, assim como seus coeficientes – como apresentado na Tabela 3.3 – é possível observar algumas das regras utilizadas pelo modelo para aumentar ou diminuir o risco de corrupção de um dado servidor público federal.

Observando o valor relativamente grande da constante *Intercept* na Tabela 3.3, assim como vendo que há diversos atributos com faixa $]-99999999999, 0.5]$ e coeficiente negativo, vê-se que o modelo opera iniciando com um alto risco de corrupção que é diminuído basicamente pela inexistência de várias características. Por exemplo, o atributo $filiados.qtd.total.filiacoes.]-99999999999, 0.5]$ de coeficiente negativo indica que a inexistência de filiação partidária diminui o risco de corrupção.

A seguir serão delineadas as principais regras obtidas pela análise dos atributos e coeficientes do modelo final, onde os efeitos foram divididos de modo geral em aumento ou diminuição do risco de corrupção.

1. **Aumentam** o risco de corrupção:

    - Possuir ou ter possuído as funções DAS 101.1, FCI 0001 e FGR 0002;
    - Ter sido envolvido em investigação da CGIE;
    - Estar ou já ter sido filiado a partido político;
    - Já ter sido penalizado em PAD com suspensão ou advertência;
    - Ter sido registrado como responsável em OS com constatação da CGU;
    - Ter pouco tempo mínimo em um mesmo cargo, segundo algumas faixas;
    - Ter participação média em vínculos com empresas maior do que aproximadamente 35%;
    - Possuir contas julgadas irregulares pelo TCU;
    - Nunca ter ocupado cargo público federal;
    - Ter se aposentado;
    - Possuir salário bruto máximo de até aproximadamente R$ 334,00;
    - Ter sido cedido pelo SUS (Sistema Único de Saúde) e possuir salário bruto máximo de até aproximadamente R$ 8.800,00.

2. **Diminuem** o risco de corrupção:

    - Ter tempo mínimo maior do que 28 anos em um mesmo cargo;
    - Possuir ou ter possuído cargo apenas em um único órgão;
    - Ser instituidor de pensão;
    - Ser celetista com menos de aproximadamente 270 dias de atividade como empregado público;
    - Ter se aposentado e possuir ou já ter possuído quatro cargos públicos.

É útil frisar ainda que, observando os coeficientes dos atributos, pode-se verificar que, das principais regras levantadas, as mais relevantes em termos de risco de corrupção – em outras palavras, as regras relacionadas com os atributos com maior valor absoluto de coeficiente – são, em ordem decrescente de influência:

1. Ter tempo mínimo maior do que 28 anos em um mesmo cargo diminui o risco de corrupção;
2. Já ter sido penalizado em PAD com advertência aumenta o risco de corrupção;
3. Já ter sido penalizado em PAD com suspensão aumenta o risco de corrupção;

4. Possuir contas julgadas irregulares pelo TCU aumenta o risco de corrupção.

Após discussão das principais regras com os especialistas da DIE, alguns comentários foram levantados de modo a racionalizar em cima do conhecimento indicado pelo modelo:

- As funções citadas como ensejadoras de aumento de risco estão muitas vezes relacionadas a chefia de unidades regionais, onde os servidores detêm muito poder decisório relativo com maior discricionariedade;
- As regras relacionadas com PAD, constatação da CGU e contas julgadas irregulares pelo TCU por si só indicam cenários de irregularidades ou impropriedades;
- A filiação a partido político está relacionada com maior influência política em decisões de interesse público a cargo dos servidores filiados;
- O registro de servidor investigado pela CGIE é realizado a partir de diversas demandas, muitas delas já relacionadas com casos de ilicitude;
- Servidor com pouco tempo mínimo em um mesmo cargo pode indicar casos de servidores que são dispostos em órgãos para trabalhos específicos ou pontuais, potencialmente relacionados a interesse ilegítmos;
- Quanto maior a participação societária de um servidor em uma empresa, maior seu interesse no resultado da mesma;
- Servidores não ocupantes de cargo público estão no governo por indicação, de caráter discricionário das autoridades;
- Servidores com tempo mínimo maior do que 28 anos e aqueles com cargo em apenas um órgão podem ser vistos como servidores mais burocratas e estabilizados;
- Um servidor aposentado que já possuiu vários cargos públicos, alcança tais cargos em grande parte por mérito próprio.

Assim, observando os comentários dos especialistas da DIE, buscando relacioná-los com as principais regras geradas pelo modelo final, vê-se que os resultados possuem razoável adequabilidade nos possíveis cenários envolvendo servidores públicos federais. Além disso, pode-se verificar alguns cenários bem específicos, provavelmente reflexão dos dados usados, onde casos pontuais são considerados relevantes, como, por exemplo, servidor cedido pelo SUS com dado salário.

4.1.6 Estudo de Casos Pontuais

Buscando validar os resultados obtidos neste trabalho, verificando os valores de risco de corrupção gerados pelo modelo final nos dados de teste (DTE), como apresentado na Seção 4.1.4, ordenou-se os resultados de forma a obter os servidores com maior e menor risco de corrupção. Com os valores para os atributos do modelo final dos dois servidores selecionados, foram analisadas as informações de cada um, de modo a atestar a validade do modelo:

- Servidor com maior risco de corrupção (aproximadamente **99%**):
    1. Com filiação a partido político;
    2. Já ocupante ativo de cargo público e em mais de um órgão diferente;
    3. Possui pouco tempo mínimo em um mesmo cargo;
    4. Foi registrado como responsável em OS com constatações da CGU;
    5. Penalizado em PAD com advertência;
    6. Penalizado em PAD com suspensão;
    7. Já cedido pelo SUS e com salário bruto máximo de até R$ 8.800,00.
- Servidor com menor risco de corrupção (aproximadamente **0%**):
    1. Nunca filiado a partido político;
    2. Possui cargo público, em apenas um único órgão e por mais do que 28 anos;
    3. Não foi registrado como responsável em nenhuma OS com constatações da CGU;
    4. Nunca foi penalizado em PAD com advertência ou suspensão;
    5. Não possui contas julgadas irregulares pelo TCU;
    6. Nunca foi investigado pela CGIE;
    7. Não foi aposentado;
    8. Nunca possuiu as funções DAS 101.1, FCI 0001 e FGR 0002.

Observando as informações dos casos pontuais, os especialistas da DIE ratificaram os resultados, atestando pela validade do modelo. Assim, consideraram úteis as informações levantadas nos dois casos pontuais estudados como indicadores de risco de corrupção de servidores públicos federais.

4.2 Implantação

Com o intuito de efetivamente atingir os objetivos pretendidos, na fase de Implantação são delineadas formas de apresentar o conhecimento gerado através do trabalho para os interessados, assim como diretrizes para colocação em uso do modelo final obtido são descritas. Usualmente, planos futuros também são elencados na fase de Implantação, no entanto, optou-se por apresentá-los no Capítulo 5 de conclusão.

4.2.1 Apresentação de Resultados

A apresentação de resultados tem como objetivo disponibilizar o conhecimento descoberto no âmbito do trabalho, de forma a possibilitar seu uso nos métodos atuais de avaliação de risco de corrupção. Dessa forma, com o compartilhamento da análise de risco de corrupção realizada procura-se contribuir com a melhoria das atividades de combate à corrupção, fornecendo insumo para análises e discussões que levam a um aumento da qualidade dos processos relacionados ao combate à corrupção. Além disso, o próprio compartilhamento dos resultados proporciona *feedback* a respeito do processo de mineração de dados utilizado, favorecendo sua otimização e o controle de possíveis inconsistências em termos de negócio.

Inicialmente, no âmbito deste projeto, a primeira análise completada e apresentada foi relacionada estritamente aos dados de filiação partidária utilizando algoritmos de classificação. Através deste trabalho específico foi construído um modelo de risco de corrupção para servidores públicos filiados que exibiu métricas de validação satisfatórias – por exemplo, precisão de 86% – e os resultados foram amplamente discutidos com os especialistas da DIE, confirmando afirmações prévias e trazendo novos pontos de vista na relação filiação-corrupção. Este estudo feito foi publicado na *Brazilian Conference on Intelligent Systems* (BRACIS) 2014, com o título de *Using Political Party Affiliation Data to Measure Civil Servants' Risk of Corruption* [55].

Como produto final do trabalho, além do conhecimento técnico em mineração de dados obtido através da consecução do projeto, e o modelo criado na Seção 3.4.2 – descrito por seus atributos e coeficientes, como visto na Tabela 3.3 – tem-se que a melhor maneira de apresentar o conhecimento gerado é delineando as regras do modelo final, nos moldes da análise feita na Seção 4.1.5. Em outras palavras, a interpretação dos atributos e seus coeficientes de forma a explicá-los em termos de cenários ou situações que os mesmos refletem.

Além disso, vê-se utilidade na visualização de casos pontuais, como o estudo feito na Seção 4.1.6, de forma a aplicar o modelo criado em casos extremos e em grupos bem

definidos, levando os resultados à discussão dos especialistas, por meio da apresentação de tais casos contrastando a resposta obtida com a esperada pelos especialistas.

Portanto, com os cenários descritos pelo modelo e os resultados de sua aplicação em determinados casos, pode-se através de relatórios estruturados e *workshops* com facilitador, levar o conhecimento obtido pelo trabalho ao alcance de todos os interessados.

4.2.2 Colocação em Uso

Com o intuito de possibilitar o uso automático das regras geradas pelo modelo obtido na Seção 3.4.2, é possível implantá-lo em um sistema computacional, de forma a habilitar sua execução em larga escala, abarcando todos os servidores públicos federais cadastrados na base de dados do SIAPE. Como tal implantação não constava como foco deste trabalho, nesta Seção apenas serão delineados os passos a serem seguidos para criar um aplicativo com regras obtidas a partir do modelo final.

Como é possível observar na Tabela 3.3, os atributos utilizados no modelo final possuem faixas específicas de valores para os quais são válidos. Dessa forma, um primeiro passo necessário à utilização do modelo é a criação de consultas SQL que gerem os dados dos atributos do modo como foram definidos, utilizando como entrada um dado CPF de um servidor público federal. Para tal, com o entendimento descrito na Seção 3.2, as definições apresentadas na Seção 3.3 e o acesso às bases de dados respectivas, é perfeitamente possível gerar tais consultas.

Na modelagem realizada na Seção 3.4.2, viu-se que o modelo final consta como obtido a partir de uma regressão logística. Portanto, como visto na Seção 2.4.2, para obtenção de uma probabilidade $Pr(Y = 1|X)$, a partir de determinados coeficientes β_i, constante α e p atributos X_i, tem-se a equação a seguir:

$$Pr(Y = 1|X) = \frac{e^{\alpha+\sum_{j=1}^{p} X_j\beta_j}}{1 + e^{\alpha+\sum_{j=1}^{p} X_j\beta_j}} \tag{4.2}$$

Assim, para o modelo final, considerando probabilidade igual a 1 para corruptos, o risco de corrupção é avaliado pela Equação 4.2, com $p = 32$ atributos e seus respectivos coeficientes β_i, sendo o valor de *Intercept* da referida tabela o equivalente a constante α da equação. A partir de tais compatibilizações, torna-se possível, portanto, a criação da função que definirá a corruptibilidade em função dos atributos levantados.

Logo, com as consultas que geram os valores dos atributos para dado servidor preparadas, juntamente com a função de probabilidade construída conforme explicitado, tem-se a possibilidade da avaliação de risco de corrupção para cada servidor público federal. Feito isso, tal processo de avaliação pode ser implementado em um aplicativo programado em diversas linguagens diferentes, fazendo uso de dados de servidores públicos federais pré-

armazenados em bases de dados específicas ou a partir da própria conexão direta com os servidores gerenciadores de bancos de dados de produção.

Capítulo 5

Conclusão e Trabalhos Futuros

Neste trabalho foi realizado o estudo e a aplicação de técnicas de mineração de dados para a criação de modelos preditivos para avaliação de risco de corrupção de servidores públicos federais, com o intuito de subsidiar o combate à corrupção realizado pela Controladoria-Geral da União (CGU). A partir das mais diversas informações disponíveis sobre servidores públicos federais e do apoio de especialistas em combate à corrupção, o modelo preditivo obtido fornece diversas medidas de risco de corrupção indicando o grau de corruptibilidade de servidores públicos federais.

Inicialmente, a partir do entendimento do negócio onde o projeto de mineração de dados encontra-se envolvido, buscou-se compreender o contexto do trabalho e o cenário de combate à corrupção no qual a CGU, e mais especificamente a DIE, está inserida. Seguindo-se com o entendimento dos dados – separando-se contextos de estudo em quatro dimensões: de Corrupção, Funcional, Política e de Vínculos Societários – diversas informações foram levantadas e categorizadas de modo a identificar aquelas relevantes para a avaliação de risco de corrupção.

Logo após, os dados levantados pelos especialistas da DIE foram pré-processados de modo a prepará-los para estarem aptos à modelagem. Além da limpeza e construção dos atributos para as etapas posteriores, foram realizadas atividades de análise de variância e correlação, separação de dados e discretização. A seleção dos atributos mais adequados para a construção de modelos foi feita em etapa envolvendo modelagem com métodos de regressão dotados de regularização. Em seguida, a construção de modelos propriamente dita foi feita, com o intuito de obter um modelo final para avaliação de risco de corrupção.

Na fase de avaliação, o modelo construído foi avaliado a partir de diversas métricas de validação – obtendo resultados satisfatórios de aproximadamente 85% de sensibilidade, 81% de precisão e 83% de acurácia – assim como foram feitos testes estatísticos corroborando a validade do modelo com nível de confiança de 95%. Em seguida, as regras geradas pelo modelo final foram analisadas, adicionando-se o estudo de casos pontuais, de modo

a subsidiar a descoberta do conhecimento obtido com o processo de mineração de dados. Finalmente, na implantação, foram delineadas formas de apresentar o conhecimento gerado através do trabalho para os interessados, assim como diretrizes para colocação em uso do modelo final obtido.

Assim, o modelo preditivo construído para avaliação de risco de corrupção de servidores públicos federais com a aplicação de técnicas de mineração de dados baseando-se no estado da arte, juntamente com um estudo detalhado do cenário no qual as informações relacionadas à corrupção se inserem, tornou possível gerar um índice de corruptibilidade para cada servidor público federal a partir de dados de diferentes fontes. Além disso, com a descoberta de conhecimento no que tange a informações sobre corruptibilidade de servidores públicos federais, levantou-se novas regras desse domínio, assim como validou-se premissas existentes, confirmando o ganho na geração de modelos preditivos.

Como planos futuros, levando em conta a continuidade deste trabalho no âmbito da DIE e buscando contribuir com o aumento da qualidade de novas atividades na mesma linha de atuação deste projeto de mineração de dados, foram levantados diversos pontos onde vislumbrou-se possibilidade de melhorias. Portanto, a seguir tais pontos são delineados, juntamente, quando for o caso, com uma breve descrição da motivação para as sugestões de trabalhos futuros:

1. realizar novo entendimento dos dados em bases de maior complexidade, como SIAPE e SICONV, buscando incluir dados não levantados no âmbito deste trabalho;
2. considerar discretizar apenas atributos selecionados, procurando manter aqueles com comportamento essencialmente linear – para tal pode-se realizar novas análises de não linearidade;
3. explorar extensamente os parâmetros dos métodos de regressão utilizados, de modo a atingir resultados ainda mais confiáveis – como, por exemplo, valores de γ no vetor de pesos adaptativos no *Adaptive Lasso*;
4. considerar ajuste mais rigoroso para informações com valores atípicos, como salário bruto e tempo de cargo, de forma a obter dados mais consistentes;
5. validar discretizações, avaliando compressão de faixas, de modo a buscar evitar faixas estreitas que favoreçam *overfitting*;
6. acrescentar atributos de cenários, onde os valores representam alguma situação específica prevista pelos especialistas, de modo a enriquecer a avaliação de risco de corrupção; e
7. buscar agrupamento de valores em atributos categóricos, com o intuito de reduzir o número de atributos e manter os dados mais concisos.

Portanto, vê-se que a construção de um modelo preditivo para análise de risco de corrupção de servidores públicos federais é de grande valia para sustentar a seleção priorizada de investigados em suspeitas de improbidade a partir de embasamento estatístico, podendo aumentar as chances de investigação de corruptos, em um trabalho feito em larga escala, impossível de ser realizado manualmente em tempo hábil considerando o atual quantitativo de aproximadamente 1 milhão servidores públicos federais ativos.

Além disso, torna possível o direcionamento de esforços de auditoria e fiscalização levando em consideração unidades com servidores de alto índice de corruptibilidade, possibilitando uma atuação mais consciente da criticidade de cada ambiente observando seus agentes alocados, otimizando o uso de recursos e pessoal da CGU e aumentando o alcance do combate à corrupção.

Este trabalho proporciona, portanto, impacto em âmbito nacional abarcando todos os estados ao analisar nível de corrupção de seus servidores públicos federais, apoiando todos os pólos regionais da CGU em termos de auditoria e fiscalização, onde com uma atuação de controle prévio há um enorme potencial de economia para os cofres públicos.

Referências

[1] Balaniuk, Remis and Bessiere, Pierre and Mazer, Emmanuel and Cobbe, Paulo. Risk based Government Audit Planning using Naïve Bayes Classifiers. *Advances in Knowledge-Based and Intelligent Information and Engineering Systems*, 2012. 26

[2] Brasil. Decreto-Lei 200 de 25 de fevereiro de 1967. Dispõe sôbre a organização da Administração Federal, estabelece diretrizes para a Reforma Administrativa e dá outras providências. *Diario Oficial da Republica Federativa do Brasil*, 27 março 1967. 61

[3] Brasil. Constituição da República Federativa do Brasil de 1988. *Diario Oficial da Republica Federativa do Brasil*, 05 outubro 1988. 46

[4] Brasil. Lei 8.112 de 11 de dezembro de 1990. Dispõe sobre o regime jurídico dos servidores públicos civis da União, das autarquias e das fundações públicas federais. *Diario Oficial da Republica Federativa do Brasil*, page 1, 19 abril 1991. 42

[5] Brasil. Lei 8.429, de 2 de junho de 1992. Dispõe sobre as sanções aplicáveis aos agentes públicos nos casos de enriquecimento ilícito no exercício de mandato, cargo, emprego ou função na administração pública direta, indireta ou fundacional e dá outras providências. *Diario Oficial da Republica Federativa do Brasil*, page 6993, 3 junho 1992. 35, 54

[6] Brasil. Lei 8.443, de 16 de julho de 1992. Dispõe sobre a Lei Orgânica do Tribunal de Contas da União e dá outras providências. *Diario Oficial da Republica Federativa do Brasil*, page 9449, 17 julho 1992. 38, 54

[7] Brasil. Lei 9.096 de 19 de setembro de 1995. Dispõe sobre partidos políticos. *Diario Oficial da Republica Federativa do Brasil*, 20 setembro 1995. 44, 45, 56

[8] Brasil. Decreto 2.271 de 7 de julho de 1997. Dispõe sobre a contratação de serviços pela Administração Pública Federal direta, autárquica e fundacional e dá outras providências. *Diario Oficial da Republica Federativa do Brasil*, page 14293, 8 julho 1997. 61

[9] Brasil. Lei 9.527 de 10 de dezembro de 1997. Altera dispositivos de algumas Leis e dá outras providências. *Diario Oficial da Republica Federativa do Brasil*, page 29421, 11 dezembro 1997. 46

[10] Brasil. Lei 10.683, de 28 de maio de 2003. Dispõe sobre a organização da Presidência da República e dos Ministérios, e dá outras providências. *Diario Oficial da Republica Federativa do Brasil*, page 2, 29 maio 2003. 32

[11] Brasil. Decreto 6.170, de 25 de julho de 2007. Dispõe sobre as normas relativas às transferências de recursos da União mediante convênios e contratos de repasse, e dá outras providências. *Diario Oficial da Republica Federativa do Brasil*, page 1, 26 julho 2007. 58

[12] Brasil. Ofício-circular nº 268/2009/SE/CGU-PR de 11 de agosto de 2009. Dispõe sobre informações sobre terceirizados. *Controladoria-Geral da União*, 11 agosto 2009. 61

[13] Brasil. Portaria CGU 516, de 15 de março de 2010. Institui o Cadastro Nacional de Empresas Inidôneas e Suspensas (CEIS) e dá outras providências. *Diario Oficial da Republica Federativa do Brasil*, 16 março 2010. 53

[14] Brasil. Decreto 7.592, de 28 de outubro de 2011. Determina a avaliação da regularidade da execução dos convênios, contratos de repasse e termos de parceria celebrados com entidades privadas sem fins lucrativos até a publicação do Decreto no 7.568, de 16 de setembro de 2011, e dá outras providências. *Diario Oficial da Republica Federativa do Brasil*, page 2, 31 outubro 2011. 55

[15] Brasil. Portaria CGU 2.379, de 30 de outubro de 2012. Institui sistemática de quantificação e registro dos benefícios do controle interno e dos prejuízos identificados. *Diario Oficial da Republica Federativa do Brasil*, 31 outubro 2012. 39

[16] Brasil. Decreto 8.109, de 17 de setembro de 2013. Aprova a Estrutura Regimental e o Quadro Demonstrativo dos Cargos em Comissão e das Funções Gratificadas da Controladoria-Geral da União e remaneja cargos em comissão. *Diario Oficial da Republica Federativa do Brasil*, page 2, 18 setembro 2013. 32, 33, 43

[17] Brasil. Lei 12.813 de 16 de maio de 2013. Dispõe sobre o conflito de interesses no exercício de cargo ou emprego do Poder Executivo federal e impedimentos posteriores ao exercício do cargo ou emprego. *Diario Oficial da Republica Federativa do Brasil*, page 1, 17 maio 2013. 34

[18] Brasil. Lei 12.846, de 1 de agosto de 2013. Dispõe sobre a responsabilização administrativa e civil de pessoas jurídicas pela prática de atos contra a administração pública, nacional ou estrangeira, e dá outras providências. *Diario Oficial da Republica Federativa do Brasil*, page 1, 2 agosto 2013. 53

[19] Brasil. Instrução Normativa RFB 1470, de 30 de maio de 2014. Dispõe sobre o Cadastro Nacional da Pessoa Jurídica (CNPJ). *Diario Oficial da Republica Federativa do Brasil*, page 23, 03 junho 2014. 50

[20] Brasil. Portaria SFC 1.161, de 30 de maio de 2014. Estabelece o padrão de elaboração e apresentação para os relatórios produzidos no âmbito da Secretaria Federal de Controle Interno. *Boletim interno da Controladoria-Geral da União*, 30 maio 2014. 39, 40

[21] Wilton de O Bussab and Pedro A Morettin. *Estatística básica*. Saraiva, 2010. 9, 14, 15, 16, 17, 20

[22] Carlos Vinícius Sarmento Silva and Célia Ghedini Ralha. Utilização de Técnicas de Mineração de Dados como Auxílio na Detecção de Cartéis em Licitações. In *WCGE - II Workshop de Computação Aplicada em Governo Eletrônico*, 2010. 26

[23] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. 2000. 5, 8, 9

[24] William S Cleveland, Susan J Devlin, and Eric Grosse. Regression by local fitting: methods, properties, and computational algorithms. *Journal of econometrics*, 37(1):87–114, 1988. 15

[25] Controladoria-Geral da União. Metodologia de mapeamento de riscos de corrupção, 2008. Available at http://www.cgu.gov.br/PrevencaodaCorrupcao/Arquivos/Metodologia.pdf. 1

[26] Tribunal de Contas da União. Instrução Normativa TCU número 71 de 28 de novembro de 2012. Dispõe sobre a instauração, a organização e o encaminhamento ao Tribunal de Contas da União dos processos de tomada de contas especial. 28 novembro 2012. 42

[27] Ramez Elmasri and S Navathe. *Database systems*. Pearson Education, 2011. 9

[28] Sergio Garcia, Julián Luengo, José Antonio Sáez, Victor Lo?pez, and Francisco Herrera. A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. *Knowledge and Data Engineering, IEEE Transactions on*, 25(4):734–750, 2013. 8, 19, 20

[29] Phillip Good. *A practitioner's guide to resampling for data analysis, data mining, and modeling*. Chapman & Hall/CRC, 2011. 13

[30] Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. On the class imbalance problem. In *Natural Computation, 2008. ICNC'08. Fourth International Conference on*, volume 4, pages 192–201. IEEE, 2008. 12

[31] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003. 63

[32] Walter W Hauck Jr and Allan Donner. Wald's test as applied to hypotheses in logit analysis. *Journal of the american statistical association*, 72(360a):851–853, 1977. 8, 21, 89

[33] Robert Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural Networks, 1989. IJCNN., International Joint Conference on*, page 593605. IEEE, 1989. 26

[34] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. 8, 15, 17, 26

[35] David W Hosmer and Stanley Lemeshow. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods*, 9(10):1043–1069, 1980. 21, 22, 89

[36] David W Hosmer, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013. 8, 14, 16, 22, 26, 75, 82, 84

[37] Keki B Irani. Multi-interval discretization of continuous-valued attributes for classification learning. 1993. 8, 19, 20

[38] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York, New York, NY, 2013. 22

[39] Nathalie Japkowicz et al. Learning from imbalanced data sets: a comparison of various strategies. 12

[40] José Matias Pereira. Reforma do Estado e transparência: estratégias de controle da corrupção no Brasil. Lisboa, Portugal, 2002. 1

[41] Randy Kerber. Chimerge: Discretization of numeric attributes. In *Proceedings of the tenth national conference on Artificial intelligence*, pages 123–128. Aaai Press, 1992. 20

[42] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, pages 1137–1145, 1995. 12, 13

[43] Yufeng Kou, Chang-Tien Lu, Sirirat Sirwongwattana, and Yo-Ping Huang. Survey of fraud detection techniques. In *Networking, sensing and control, 2004 IEEE international conference on*, volume 2, pages 749–754. IEEE, 2004. 26

[44] Max Kuhn. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008. 82

[45] Lukasz A Kurgan and Krzysztof J Cios. Caim discretization algorithm. *Knowledge and Data Engineering, IEEE Transactions on*, 16(2):145–153, 2004. 8, 19

[46] Kathryn Blackmond Laskey. MEBN: A language for first-order bayesian knowledge bases. *Artificial Intelligence*, 172(2-3):140–178, February 2008. 27

[47] Huan Liu and Rudy Setiono. Feature selection via discretization. *IEEE Transactions on Knowledge & Data Engineering*, (4):642–645, 1997. 20

[48] Shuangge Ma and Jian Huang. Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics*, 9(5):392–403, 2008. 26

[49] R. W. M. Nelder, J. A. e Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society*, 135:370–384, 1972. 14, 15

[50] EWT Ngai, Yong Hu, YH Wong, Yijun Chen, and Xin Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3):559–569, 2011. 26

[51] Joseph O Ogutu, Torben Schulz-Streeck, and Hans-Peter Piepho. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. In *BMC proceedings*, volume 6, page S10. BioMed Central Ltd, 2012. 26

[52] Thomas Oommen, Laurie G Baise, and Richard M Vogel. Sampling bias and class imbalance in maximum-likelihood logistic regression. *Mathematical Geosciences*, 43(1):99–120, 2011. 12

[53] Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010. 26

[54] GIANNIS Potamitis and DR SANDRA SAMPAIO. Design and implementation of a fraud detection expert system using ontology-based techniques. *A dissertation submitted to the University of Manchester for the degree of Master of Science in the Faculty of Engineering and Physical Sciences*, 2013. 26

[55] Ricardo Silva Carvalho, Rommel Novaes Carvalho, Marcelo Ladeira, Fernando Mendes Monteiro and Gilson Liborio Mendes. Using political party affiliation data to measure civil servants' risk of corruption. In *2014 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 166–171. IEEE, 2014. 95

[56] Rommel Novaes Carvalho. *Representation and Modeling Methodology*. PhD thesis, George Mason University, 2011. 27

[57] Rommel Novaes Carvalho, Shou Matsumoto, Kathryn B. Laskey, Paulo C. G. Costa, Marcelo Ladeira, and Laécio L. Santos. Probabilistic ontology and knowledge fusion for procurement fraud detection in brazil. In *Uncertainty Reasoning for the Semantic Web II*, pages 19–40. Springer, 2013. 27

[58] Rommel Novaes Carvalho, Leonardo Sales, Henrique Rocha and Gilson Liborio Mendes. Using Bayesian Networks to Identify and Prevent Split Purchases in Brazil. In *BMAW*, Quebec City, Canada, 2014. 26

[59] Leonardo Sales. Risk prevention in brazilian government contracts using credit scoring. *Interdisciplinary Insights on Fraud*, 2014. 58

[60] Giovanni Seni and John F Elder. Ensemble methods in data mining: improving accuracy through combining predictions. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2(1):1–126, 2010. 26

[61] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *AI 2006: Advances in Artificial Intelligence*, pages 1015–1021. Springer, 2006. 23

[62] Daniel B Suits. Use of dummy variables in regression equations. *Journal of the American Statistical Association*, 52(280):548–551, 1957. 65

[63] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, page 37, 2014. 25, 26

[64] Francis EH Tay and Lixinang Shen. A modified chi2 algorithm for discretization. *Knowledge and Data Engineering, IEEE Transactions on*, 14(3):666–670, 2002. 8, 19, 20

[65] Richard Taylor. Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 6(1):35–39, 1990. 78

[66] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 8, 15, 17, 18, 26

[67] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2008. 8, 9, 13, 14, 15, 17, 18, 19, 26, 79

[68] Dongsheng Tu and J Shao. The jackknife and bootstrap. *Springer Series in Statistics, New York*, 85:486–492, 1995. 8, 12, 13, 83, 84

[69] Stéphane Tufféry. *Data mining and statistics for decision making*. John Wiley & Sons, 2011. 8, 14, 15, 19, 26, 76, 79

[70] Gary M Weiss and Foster Provost. Learning when training data are costly: the effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, pages 315–354, 2003. 12

[71] Sholom M. Weiss. *Predictive data mining: a practical guide*. Morgan Kaufmann, 1998. 8, 9

[72] Mohammed J Zaki and Wagner Meira. *Data mining and analysis: fundamental concepts and algorithms*. 2013. 9

[73] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006. 8, 15, 18, 19, 26