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

Modelos Preditivos para Avaliação de Risco de Corrupção de Servidores Públicos do Distrito Federal

Marcelo Oliveira Vasconcelos

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

Orientador
Prof. Dr. Ricardo Matos Chaim

Brasília
2021

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

VV331m

Vasconcelos, Marcelo Oliveira
Modelos Preditivos para Avaliação de Risco de Corrupção deServidores Públicos do Distrito Federal / Marcelo Oliveira Vasconcelos; orientador Ricardo Matos Chaim. -- Brasília, 2021.
72 p.

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

1. Corrupção. 2. Servidor Público. 3. Mineração de dados. 4. Riscos. 5. CRISP-DM. I. Chaim, Ricardo Matos, orient. II. Título.

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

Modelos Preditivos para Avaliação de Risco de Corrupção de Servidores Públicos do Distrito Federal

Marcelo Oliveira Vasconcelos

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

Orientador
Prof. Dr. Ricardo Matos Chaim

Prof. Dr. João Carlos Félix Souza
FT/UnB

Prof. Dr. Rosalvo Ermes Streit
UCB

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

Brasília, 30 junho de 2021.

Dedicatória

Agradeço a Deus, fonte criadora e inteligência suprema, que é o princípio de tudo e causa primária de todas as coisas.

Aos meus pais; a minha esposa, Rebeka, e filhos, Isabelle e Lorenzo, não só pela paciência e compreensão nas intermináveis horas de dedicação e ausência que foram necessárias durante a execução desse projeto. Mas principalmente pelo incentivo e apoio da minha amada esposa, sempre me motivando a tentar ir além. E também aos meus filhos, pelo incentivo de poder ser uma inspiração para eles.

Aos diversos colegas de trabalho que incentivaram e auxiliaram nessa atividade. Em especial, ao meu chefe, Flávio, por me apoiar desde o início, pela confiança em minha capacidade de realizar essa tarefa, e por me dar condições de dedicar tempo a esse trabalho. A equipe do TCDF – Rômulo, Fred, Ligu, Vilcemar, Júlio, Alcuri, Ton, Adinor, Luís, Cláudio, Angelo – que sempre foram solidários e que são motivo de orgulho para mim por poder fazer parte desse grupo de profissionais seleto e dedicados. Espero poder retribuir aplicando o que aprendi e, principalmente, procurando ser um profissional melhor para a instituição.

“Agradeço todas as dificuldades que enfrentei; não fosse por elas, eu não teria saído do lugar. As facilidades nos impedem de caminhar. Mesmo as críticas nos auxiliam muito.” Chico Xavier

Resumo

A pesquisa tratou de elaboração de um modelo preditivo para avaliação de risco de corrupção de servidores públicos do Distrito Federal, nos termos da Lei nº 8.429/92, com auxílio da literatura acadêmica para identificação de atributos e definição de algoritmos de aprendizagem de máquina para aplicação na fiscalização do Tribunal de Contas do Distrito Federal (TCDF), para isso levantou-se a literatura acadêmica dos fatores de risco e das técnicas de aprendizagem de máquina relacionadas à corrupção e com auxílio de especialistas, seguindo o método CRISP-DM como referência, iniciou-se o processo de mineração com integração de oito bases de dados com seleção de atributos, limpeza, transformação, análise de variância e de correlação, separação de dados e modelagem. O algoritmo utilizado foi regressão logística e uma das dificuldades da investigação foi o desbalanceamento extremo de classes a razão de 1:707 ou em termos percentuais 0,14% da classe de interesse em relação a população. Para solução foram utilizadas duas possíveis abordagens, balanceamento com técnicas de reamostragem com uso de *synthetic minority oversampling technique* SMOTE e aplicação de algoritmos com características específicas de parametrização para obter os padrões desejados da classe minoritária de forma a evitar viés da classe dominante. O melhor resultado de modelagem ocorreu com a aplicação da técnica de pesos gerando valor de área sobre a curva ROC de 0,7, definição de sessenta e oito atributos e seus respectivos coeficientes que correspondem aos fatores de risco de corrupção. O resultado dessa pesquisa foi a identificação dos fatores de risco de corrupção dos servidores do GDF a partir do modelo gerado para que com esses parâmetros possa auxiliar na definição de planejamento de fiscalização do TCDF com otimização de recursos (pessoal e equipamentos) e foco nas atividades de maior risco para Administração Pública. Como a melhoria futura a este trabalho, será a inclusão de atributos do sistema de compras governamentais do GDF (Ecompras) que está previsto para operar em meados de 2021.

Palavras-chave: Corrupção, Servidor Público, Mineração de dados, Riscos, CRISP-DM

Abstract

The research aims to create a predictive model for the assessment of the risk of corruption of public servants in the Federal District, under the terms of Law No. 8,429/92, by using academic literature to identify attributes and define machine learning algorithms to be applied in the inspection of the Federal District Court of Auditors (TCDF), for this purpose the academic literature on risk factors and machine learning techniques related to corruption was raised and with the help of specialists and following the CRISP-DM method as a reference, the data mining process began with the integration of eight databases with the features selection, creation of a dataset, cleaning, transforming, and modeling. The algorithm used was logistic regression, and one of the difficulties of the investigation was the extreme imbalance of classes at a ratio of 1:707 or, in percentage terms, 0.14% of the interest class to the population. For solution, two possible approaches were used, balancing with resampling techniques using *synthetic minority oversampling technique* SMOTE or applying algorithms with specific parameterization characteristics to obtain the desired standards of the minority class without generating bias from the dominant class. Finally, the best modeling result was obtained by applying the weights technique, generating an area value on the ROC curve of 0.7, defining sixty-eight features and their respective coefficients that correspond to the risk factors for corruption. The result of this research aimed to identify the risk factors of corruption of the civil servants of the GDF so that these parameters can assist in the definition of overseen planning of the TCDF with optimization of resources (personnel and equipment) and focus on the activities of greatest risk for Public Administration, that is, cases that have a high probability of occurrence and a high financial or social impact. A future improvement to this work includes features of the GDF government procurement system (Ecompras) that is expected to operate in 2021.

Keywords: Corruption, Civil Servant, Data Mining, Risk, CRISP-DM

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

1.1 Contextualização

A corrupção é um problema comum em países em desenvolvimento [1] que acarreta aumento do custo do serviço público, prejudica o crescimento econômico [2] e dificulta a condução dos negócios privados.

O conceito de corrupção mais aceito na literatura internacional foi dado pela Transparência Internacional "Corrupção é o abuso do poder confiado para ganhos privados."[3].

Apesar desse conceito nortear os trabalhos referenciados nessa dissertação, será utilizado o conceito descrito na Lei nº 8.429, de 2 de julho de 1992, que define corrupção como ato de improbidade que, sob influência ou não do cargo provoque enriquecimento ilícito, cause ou não lesão ao erário ou viole princípios da Administração Pública [4].

Um dos casos de corrupção mais marcantes no Brasil foi descoberto com uma investigação de uma rede de postos de combustíveis e lava a jato pertencentes a uma organização criminosa. Essa apuração avançou e expandiu-se para outras organizações criminosas tornando-se a maior investigação de corrupção e lavagem de dinheiro do país. Estimam-se desvios de recursos da empresa Petrobrás da ordem de bilhões de reais [5].

Segundo Sérgio Moro, as apurações dessa operação, conhecida como "Lava Jato", identificaram o envolvimento de políticos e seus partidos, de empresários donos de empreiteiras e de funcionários públicos da Petrobrás indicados por partidos políticos para compor a alta gestão da empresa [6].

Diversas pessoas foram julgadas e condenadas por crimes de colarinho branco – propina e lavagem de dinheiro. Esse tipo de crime é considerado algo difícil de se descobrir, provar e punir, porque, em geral, se caracteriza pela existência de segredo entre as partes, de pessoas investidas de alto poder no escalão de governo e um certo grau de sofisticação para qual os órgãos de investigação não estão preparados [6].

A corrupção, como crime isolado, existe em todo o mundo, mas a corrupção sistemática – pagamento de propinas como regra do jogo – não é algo comum, e representa grave degeneração da função pública e privada, especialmente em nações democráticas [7].

O custo da corrupção sistemática é expressivo. Parte desse custo é a propina que se acrescenta ao valor do contrato onerando o orçamento público. Outra parte, para administração pública, é a má gestão dos recursos públicos que gera baixa qualidade da prestação do serviço público e, no caso das empresas governamentais, também se acrescenta a desvalorização dos ativos e perdas dos investimentos nacionais e internacionais [7].

Outro aspecto da corrupção é a inadequada tomada de decisão de investimentos das entidades públicas e privadas que, por meio dos dirigentes corruptos, atuando de forma a maximizar a propina, estabelecem escolhas sem considerar a economicidade e a melhor estratégia para a entidade.

Não existe consenso na melhor forma de reduzir a corrupção [1], no entanto, duas formas de agir podem ser empregadas. Atuação *a posteriori*, ou seja, após identificado o dano agir para recuperação dos valores; ou de forma preventiva, agindo de forma a coibir a ocorrência dos fatos [8].

A busca pela recuperação dos recursos fraudados não é ação eficiente. Segundo o Relatório das Nações, elaborado pela *Association of Certified Fraud Examiners* [9], das fraudes que ocorreram em 125 países examinados, em cerca de 15% dos casos houve a recuperação total dos valores e 32% de forma parcial, restando 51% sem nenhum retorno dos valores identificados.

No entanto, a prevenção mostra-se como a forma mais eficiente e econômica de coibir a corrupção, pois evita um processo moroso de recuperação de valores e permite ação baseada em riscos. O estudo de Gans-Morse [10] aponta que uma política bem sucedida de combate a corrupção é baseada em monitoramento, por auditorias anticorrupção e *e-governance*, i.e., governança eletrônica.

A governança eletrônica foi definida por Gans-Morse [10] como uso da tecnologia da informação para fornecer serviços governamentais. Este serviço permite ampliar a transparência e está cada vez mais difundido, permitindo o monitoramento pela sociedade e dificultando a possibilidade de atos de corrupção.

Atualmente, os órgãos de fiscalização no Brasil apresentam ao seu dispor diversas bases de dados que permitem exames minuciosos e identificação de padrões de desvios de conduta de pessoas ou empresas que podem ser mapeados em tipologias e utilizados para investigações.

Os tribunais de contas no Brasil são órgãos técnicos e independentes que fazem parte da estrutura do Poder Legislativo cuja especialidade é fiscalizar, sob aspecto técnico, a gestão financeira e orçamentária e contribuir com o aperfeiçoamento da Administração Pública em benefício da sociedade. Para esse fim, o emprego da tecnologia da informação permite impulsionar e alavancar o controle externo por meio de tratamento e análise de dados [11].

A título de exemplo, o Tribunal de Contas da União – TCU relatou casos de sucesso na aplicação de técnicas de mineração de dados e métodos de identificação de fraudes por detecção de anomalias estatísticas no sistema InfoSAS para tratamento de dados do SUS [12].

Em entrevista publicada na Revista TCU, Timothy Persons, cientista chefe do *Government Accountability Office – GAO* dos Estados Unidos [13], entende que "a fraude e a corrupção são análises efetivas de custo benefício feitas por pessoas que podem cair na tentação de se beneficiar à custa dos demais". Persons acrescenta que a proliferação de abordagens antifraude eficazes de análises de dados aumentaram o risco e o custo de alguém ser descoberto fazendo coisas nefastas e de ser processado, consequentemente reduzindo de forma eficiente o comportamento fraudulento.

Persons expressa que o Tribunal de Contas da União poderá se beneficiar muito das abordagens analíticas de fraudes, tais como a vinculação de software para rede social/rastreamento de fundos como análises, a análise geoespacial, a extração de texto de grandes conjuntos de dados para fins de construção de sentido, entre outras [13].

Considerando estudos realizados sobre corrupção e consulta a especialistas do tema, esta pesquisa buscou o estudo e aplicação de técnicas de mineração de dados para criação de modelo preditivo para avaliação de risco de corruptibilidade de servidores públicos do Distrito Federal. A identificação dos atributos relevantes pelo processo de mineração servirá em trabalho futuro para subsidiar a fiscalização para o combate a corrupção pelo Tribunal de Contas do Distrito Federal – TCDF.

O presente tema trata de assunto voltado a atividade de inteligência, pois conforme definição da Lei nº 9.883/1999 [14], consiste em busca e coleta de dados para processamento e produção de informação para estabelecer avaliação de risco que possa orientar a tomada de decisão pelo decisor, possibilitando minimizar as incertezas [15] e ação nos assuntos de interesse da sociedade e do Estado.

1.2 Descrição do problema

O contexto das ações de fiscalização do TCDF é a ausência de critério que considere riscos de corrupção de servidores, também não há estudo no âmbito do Governo do Distrito Federal que tenha contribuído para elaboração de um modelo preditivo de corrupção de servidores públicos.

Como a prática de corrupção está relacionada a ação de servidores públicos [16][4][17], considera-se necessário priorizar as atividades de combate à corrupção. Nesse contexto, entende-se apropriado o uso de métodos automatizados para enfrentamento do problema.

Este cenário leva à oportunidade de se investigar as características do comportamento de corrupção no setor público e dos casos comprovados de corrupção para estabelecer modelo preditivo de corrupção de servidores públicos que possa ser utilizado para otimizar a fiscalização da instituição e em última instância possibilitar gerar melhor benefício à sociedade na prestação do serviço público.

1.3 Justificativa

Diversas ações de fiscalização de corrupção são iniciadas a partir de denúncias recebidas. Este modo operacional não é caracterizado por ação de planejamento de fiscalização, nesse aspecto, mostra-se adequado o levantamento de características de corrupção para aplicação de fiscalização com base em riscos de corrupção.

Para isso, a análise de dados mostra-se como uma abordagem capaz de permitir a otimização dos recursos do Tribunal (pessoal e equipamentos) para melhor eficiência de sua atividade [18] e se associado a técnicas de gerenciamento de riscos permitir melhor estabelecer uma classificação de risco nas fiscalizações e direcionar os recursos do TCDF aos objetos que tenham maior risco a Administração Pública, ou seja, casos que tenham alta probabilidade de ocorrência e alto impacto financeiro ou social [15].

A aplicação de um modelo voltado a área de inteligência em um órgão de fiscalização busca otimizar a atuação do órgão de controle. Enquanto a área de inteligência busca obter informação sem gerar produção de provas, a fiscalização depende essencialmente de evidências para formação de juízo. No entanto, o auxílio da área de inteligência permite tornar ágil a atuação da fiscalização ao direcioná-la para os casos de maior risco de danos para administração pública e para sociedade.

Nesse contexto, a pesquisa mostra-se relevante pois ao criar modelo preditivo para análise de risco de corrupção de servidores públicos, com uso de técnicas de mineração de dados, permite gerar subsídio para priorizar as ações de fiscalização com base em estatística aplicada em base de dados possibilitando tornar mais eficaz a atuação do Tribunal na prevenção de corrupção e redução de gastos públicos.

1.4 Objetivos

Nesta seção estão descritos o objetivo geral e os objetivos específicos.

1.4.1 Objetivo Geral

Com auxílio da literatura acadêmica para identificação de atributos e definição de algoritmos de aprendizagem de máquina, elaborar um modelo preditivo para avaliação de risco de corrupção de servidores públicos do Distrito Federal, nos termos da Lei nº 8.429/92, para aplicação na fiscalização do Tribunal de Contas do Distrito Federal com validação de especialistas da área de negócio.

1.4.2 Objetivos Específicos

Para alcançar o objetivo geral desta dissertação, os seguintes objetivos específicos foram buscados:

I – Identificar os fatores de riscos relativos à corrupção de servidores públicos;

II – Identificar as técnicas de mineração de dados para o contexto de corrupção;

III – Elaborar um modelo preditivo a partir de mineração de dados em bases de dados governamentais do DF e da União afetos aos servidores do DF, com os resultados dos objetivos específicos anteriores e com auxílio de especialistas em corrupção do Tribunal de Contas do DF na interpretação e criação do modelo;

IV – Validar os resultados com os especialistas do TCDF.

1.5 Contribuições

A principal contribuição desta pesquisa é gerar um modelo preditivo que possa indicar os principais fatores de risco para auxiliar a atuação do planejamento das atividades de fiscalização do Tribunal de Contas do DF.

Esses fatores possibilitam contribuir para priorização de ações de auditoria e de inspeções na Administração Pública do DF conforme criticidade de cada caso, com relação a um servidor, grupo de servidores, ou processo, quando identificável, que podem representar risco considerável à administração ou à prestação do serviço público a sociedade.

Outro aspecto é a possibilidade de descoberta de conhecimento por padrões ainda não identificados, mas que possam ser avaliados pelos especialistas em casos de corrupção.

Uma forma de atuação do TCDF é emitir decisões, normas e orientações de caráter obrigatório à administração pública do GDF. Essa prerrogativa do tribunal permite aplicação de controles capazes de restringir ou inibir atos de corrupção em atividades ou processos que possam ser identificados da análise dos padrões obtidas da mineração de dados.

Nesse contexto, a identificação de situações de risco de danos ao erário ou à disponibilização dos serviços públicos sujeita-se à orientação do Tribunal de Contas, inclusive fundamentada na Lei de Integridade da Administração Pública do DF [19] e em normas de boas práticas como a ABNT ISO 37001-2017 [20] que trata de sistemas de gestão antissuborno.

Como o resultado dessa pesquisa se caracterizou por informação de inteligência, seu produto permite contribuir para o combate a corrupção exercido pelo TCDF em ações de fiscalização.

1.6 Estrutura dos Capítulos

A pesquisa foi estruturada em seis capítulos, sendo que o primeiro apresentou a introdução, o Capítulo 2 descreveu os procedimentos metodológicos aplicada nessa pesquisa. O Capítulo 3 abordou a fundamentação teórica dos principais assuntos que representam referência teórico de corrupção de servidor público e a revisão de literatura das técnicas de mineração aplicadas a corrupção e fraude. A solução proposta foi delineada no Capítulo 4. E os Resultados no Capítulo 5. E por fim, no capítulo 6, expôs-se a conclusão e trabalhos futuros.

2. Procedimentos Metodológicos

Para o alcance do objetivo geral e de cada um dos objetivos específicos propostos nesta pesquisa, apresenta-se neste capítulo os procedimentos metodológicos em cinco seções, a saber: tipo de pesquisa, universo da pesquisa, bases de dados, mineração de dados e validação por especialistas.

2.1. Tipo de pesquisa

Quanto à natureza, este trabalho pode ser classificado como Aplicado, pois visou extrair conhecimento de padrões por mineração de dados para gerar informação de fatores de risco de corrupção de servidores públicos do Distrito Federal e permitir melhorias em procedimentos de fiscalização do Tribunal de Contas do Distrito Federal.

Quanto ao objetivo, a pesquisa classifica-se como Pesquisa Exploratória, pois envolve levantamento bibliográfico para definição de atributos, técnicas de mineração e entrevistas com auditores que atuam na área de fiscalização com experiência no problema de pesquisa. Segundo Gil [21], a pesquisa exploratória proporciona maior familiaridade com o problema para torná-lo mais claro e preciso.

Quanto à abordagem, caracteriza-se como pesquisa qualitativa-quantitativa [22]. Sob o ponto de vista quantitativo, utilizou-se de técnicas e ferramentas estatísticas, algoritmo de aprendizagem de máquina – regressão logística, como meio de análise dos dados na pesquisa. Enquanto sob o ponto de vista qualitativo, empregou-se entrevistas e interpretação dos resultados da regressão logística sob enfoque de corrupção para identificação de fatores de risco que pudessem contribuir na fiscalização do TCDF.

2.2. Universo da Pesquisa

Em relação ao universo de pesquisa, a pesquisa é censitária, pois engloba todos os servidores e pensionistas do Governo do Distrito Federal, totalizando 303.036 (trezentos e três mil e trinta e seis) em janeiro de 2020.

Para essa população da pesquisa foram agregados diversos atributos de dados que pudessem contribuir na investigação dos fatores de risco de corrupção de servidores públicos, a partir de oito bases de dados que serão descritas na próxima seção.

2.3. Bases de Dados

A definição de bases de dados que foram empregadas para investigação ocorreu segundo alguns passos. Inicialmente pela identificação de atributos pela execução do objetivo específico 1 – Identificar os fatores de riscos relativos à corrupção de servidores públicos, para isso, elaborou-se estudo meta-analítico com adaptações para identificação da produção científica e formação do arcabouço relativo à corrupção em diversos países (seção 3.2).

O segundo passo foi a identificação das bases de dados que poderiam apresentar os atributos encontrados na pesquisa de literatura do item anterior.

Desse exame, verificou-se que poucos atributos não foram possíveis de se obter para a presente investigação, conforme descrito na seção 4.2 – Entendimento dos dados. No entanto, não representaram óbice à execução e obtenção de resultados adequados nessa pesquisa.

A partir das bases de dados disponíveis no TCDF foi possível estabelecer um conjunto de dados consolidados com os atributos disponíveis para o processo de mineração de dados.

Esses atributos foram classificados em quatro áreas de conhecimento denominadas dimensões (corrupção, funcional, política e vínculos societários), assim como no trabalho de Carvalho [48], e foram obtidas a partir da integração de oito bases de dados do Governo Federal e do Distrito Federal representadas na figura a seguir.

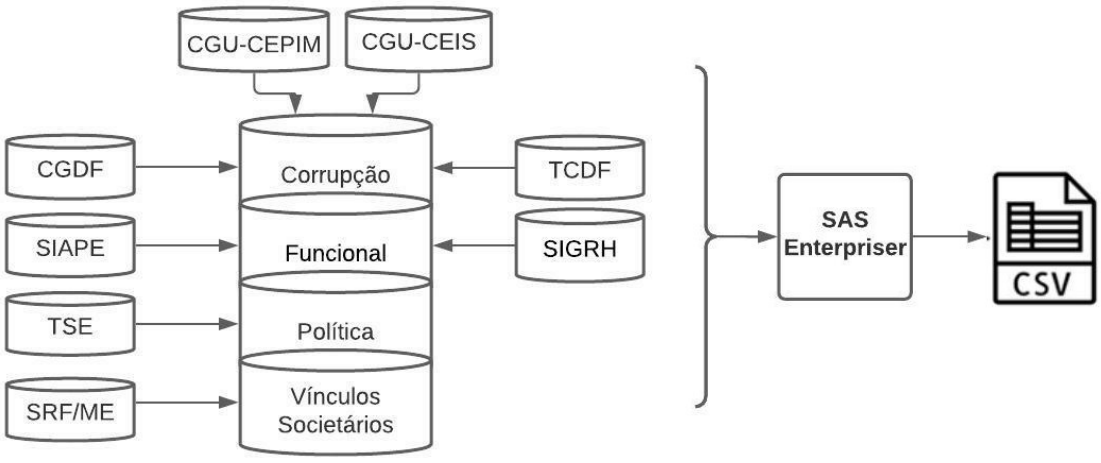


Figura 2.1: Processo de ETL (*Extract, transform and Load* - extração, transformação e carga)

O conjunto de dados consolidado foi criado após o processo de ETL com a integração das seguintes bases de dados:

- ✓ Corregedoria Geral do Distrito Federal (Portal da Transparência DF);

- ✓ SIGRH - Sistema Integrado de Gestão de Recursos Humanos;
- ✓ SIAPE - Sistema Integrado de Administração de Recursos Humanos;
- ✓ TCDF com informações de inabilitados para o exercício de cargo em comissão ou função de confiança no âmbito da Administração Pública do Distrito Federal;
- ✓ Entidades Privadas sem Fins Lucrativos Impedidas de contratar com a Administração Pública (CEPIM) Controladoria Geral da União;
- ✓ Cadastro de Empresas Inidôneas e Suspensas (CEIS) Controladoria Geral da União;
- ✓ Tribunal Superior Eleitoral (Dados eleitorais);
- ✓ Secretaria da Receita Federal (dados cadastrais de pessoa física e jurídica)

Cabe destacar que os dados presentes nessas bases de dados representam dados públicos que ao longo da investigação passaram a ser publicados pelos órgãos detentores dessa informação com base na Lei de Acesso à Informação com devida proteção a informação pessoal sensível, garantida pelo direito constitucional de respeito à intimidade, vida privada, honra e imagem das pessoas.

Das instituições governamentais descritas nesse rol de base de dados somente a Secretaria da Receita Federal apresentou maior resistência a divulgação de dados devido a possibilidade de violação de sigilo fiscal, no entanto, após discussão do tema no âmbito dos Processo nºs 168530.04455/2018-31 e 16853.008858/2017-78 de Lei de Acesso à Informação da Controladoria Geral da União foi determinado à Secretaria da Receita Federal a divulgação dos dados com descaracterização do CPF dos sócios das empresas.

Desse modo, a presente pesquisa trata de dados públicos, que podem, em regra, ser obtidos diretamente pelas consultas aos sítios das instituições públicas, sítios de dados abertos ou a pedido, com fundamento na Lei de Acesso à Informação (Lei nº 12.527/11).

A partir desse rol de bases de dados foi elaborado a consolidação dos dados em apenas um arquivo com os diversos atributos relacionados a pesquisa.

2.4. Mineração de Dados

O processo de mineração de dados foi realizado seguindo os passos do modelo de referência CRISP-DM (secção 3.1), e esse processo está diretamente relacionado ao objetivo específico 3, que é a elaboração do modelo preditivo.

Para execução deste objetivo específico, foi utilizado as bases de dados descritas na secção anterior que utilizou os resultados dos objetivos específicos 1 (secção 3.2) e 2 (secção 3.3), levantamento de atributos e técnica de mineração de dados a partir de pesquisa na literatura.

A investigação de técnicas de mineração também identificou a necessidade de abordar o problema de desbalanceamento de classes ou balanceamento de dados (secção 3.4), que é uma característica de dados no contexto de fraude e corrupção em que a classe de interesse representa baixo percentual da população examinada.

A necessidade de solucionar essa questão se deve à tendência de os algoritmos de aprendizagem de máquina predizerem valores para as classes majoritárias e, portanto, desprezarem a classe minoritária, que é a classe de interesse em que os padrões e tendências devem ser reconhecidos por estes algoritmos para posterior identificação dos fatores de risco objeto dessa pesquisa.

De forma a evitar o viés dos algoritmos para a classe majoritária, duas abordagens foram obtidas da literatura: a primeira, com tratamento de dados de forma a equilibrar as classes para treinamento do algoritmo, e a segunda, a partir da utilização de algoritmos otimizados para uso com classes desbalanceadas.

Segundo a literatura, a primeira abordagem apresenta como expoente a técnica *Synthetic Minority Oversampling Technique – SMOTE* [23], para balanceamento de dados e foram apresentados nessa pesquisa os resultados de oito variações dessa técnica (tabela 4.1).

Quanto à segunda abordagem, algoritmo otimizado, aplicou-se duas técnicas: uma com atribuição de pesos na proporção de casos da classe minoritária em relação a majoritária e outra uma implementação heurística de melhores práticas, ambas disponíveis na biblioteca *Scikit-Learn* implementadas em *Python* (tabela 4.1).

As duas abordagens foram executadas e avaliadas segundo desempenho de preditor binário pela área sobre a curva de característica de operação do receptor (*receiver operating characteristic* – ROC, definido por Mandrekar [24] (secção 3.6), que é um gráfico de

sensibilidade em relação à especificidade de um teste e a medida de área sobre essa curva expressa a medida de desempenho do modelo.

Depois de selecionado o melhor resultado, implementou-se o modelo para o algoritmo de regressão logística (LOGIT) e foram obtidos os coeficientes da regressão logística (tabela 4.3).

A expressão de intensidade dos fatores de risco de corrupção corresponde a exponencial neperiana de cada coeficiente da regressão logística, também conhecido como *Odds Ratio* (secção 5.3).

Por fim, esses resultados foram discutidos com os especialistas de cada área das dimensões definidas de forma a validar ou não os resultados (secções 5.3 e 5.4).

Quanto às ferramentas utilizadas na investigação, foi empregada a ferramenta *SAS Enterprise Guide* para o tratamento de dados e a IDE Anaconda com uso de Python e R Studio para aplicação dos algoritmos de mineração de dados.

2.5. Validação por especialistas

Esta secção engloba o objetivo específico 4, que trata da validação dos resultados com os especialistas de cada área de atuação.

Os especialistas que participaram da pesquisa são auditores com conhecimento de negócio e/ou dados que atuam na fiscalização em diversos setores:

- ✓ dois na área de legislação de pessoal;
- ✓ um auditor com experiência na fiscalização de prestação de contas partidárias junto ao Tribunal Superior Eleitoral;
- ✓ três auditores do Núcleo de Informações Estratégicas; e
- ✓ três auditores em atividades de análise de licitações com experiência em casos de corrupção.

A condução dos trabalhos foi essencialmente com entrevistas informais com periodicidade e profundidade a depender da necessidade do caso, seguindo como roteiro de atividades os seguintes questionamentos por área de especialidade de cada auditor com fins de:

1. Auxílio na construção de atributos com vista a possibilitar transformar dados de sistemas essencialmente transacionais ou gerenciais em atributo com sentido para área de negócio da pesquisa, corrupção;

2. Auxílio na integração de bases que dependem de conhecimento de negócio da área específica de cada dimensão;
3. Auxílio no tratamento de dados que poderiam ser para limpeza de dados (exclusão de outliers, ausência de valores ou valores incorretos), ou construção de atributos;
4. Auxílio na interpretação dos resultados de preditores da regressão logística e validação quanto a adequação do enquadramento como fator de risco de corrupção.

As diversas sugestões desse corpo técnico foram aplicadas nas quatro dimensões definidas e, ainda, na escolha da técnica de aprendizado de máquina, conforme descrito a seguir.

Com os resultados do objetivo específico 2 – Identificar as técnicas de mineração de dados para o contexto de corrupção (seções 3.3 e 5.2), elaborado com base no estudo meta-analítico para identificação da produção científica relativa às técnicas de mineração aplicadas à corrupção e fraude.

Dessa investigação, discutiu-se com os especialistas os resultados das técnicas de mineração e foi definido a aplicação da regressão logística nessa pesquisa, pois a variável de interesse é dicotômica (corrupção ou não corrupção); o algoritmo apresenta fácil entendimento de uso e interpretação de resultado em relação a outras técnicas como aprendizado aprofundo (*deep learning*) e, ainda, como não existe maturidade da instituição no uso de inteligência artificial, como no caso de outros órgãos da Administração Pública, poderia dificultar a implementação e uso das técnicas com menor transparência.

3. Fundamentação Teórica

Este capítulo apresenta a revisão de literatura que buscou resumir as discussões dos trabalhos da literatura a respeito de corrupção para embasamento da presente pesquisa e as abordagens de mineração de dados aplicáveis ao tema de corrupção.

3.1. Modelo de referência CRISP-DM

Fayyad [25], em 1996, explica que o termo KDD, *knowledge discovery in databases*, foi criado em 1989, para definir um processo amplo de encontrar conhecimento em dados e destaca os desafios da mineração de dados em reconhecer padrões, do aprendizado de máquina e algoritmos que deveriam ser carregados na memória do computador.

Segundo Fayyad, em 1996, os equipamentos necessários para essa aplicação não eram disponíveis de forma ampla para empresas e nem para a comunidade científica, mas com a redução substancial dos custos de equipamentos (inclusive memória) a mineração de dados pode assumir amplo uso comercial e acadêmico.

Atualmente, dois modelos de referência para processo de mineração de dados estão bem difundidos: SEMMA e CRISP-DM.

Segundo o instituto SAS, o acrônimo SEMMA é o processo de mineração de dados oriundo das iniciais de **S**ample, **E**xplorer, **M**odify, **M**odel e **A**ssess [26].

O modelo de referência CRISP-DM, corresponde a abreviação de **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining e é um processo desenvolvido pelo esforço de um consórcio inicialmente composto pelas instituições NCR, SPSS e DaimlerChrysler [27].

Um estudo comparativo [28] dos dois modelos de referência concluiu que os dois são implementações do processo KDD, descrito por Fayyad [25], e o modelo CRISP-DM é mais completo do que o modelo SEMMA.

Desse modo, optou-se pela utilização do modelo de referência CRISP-DM nesse trabalho.

O modelo CRISP-DM é um ciclo de vida que consiste em seis fases, detalhadas a seguir [27].

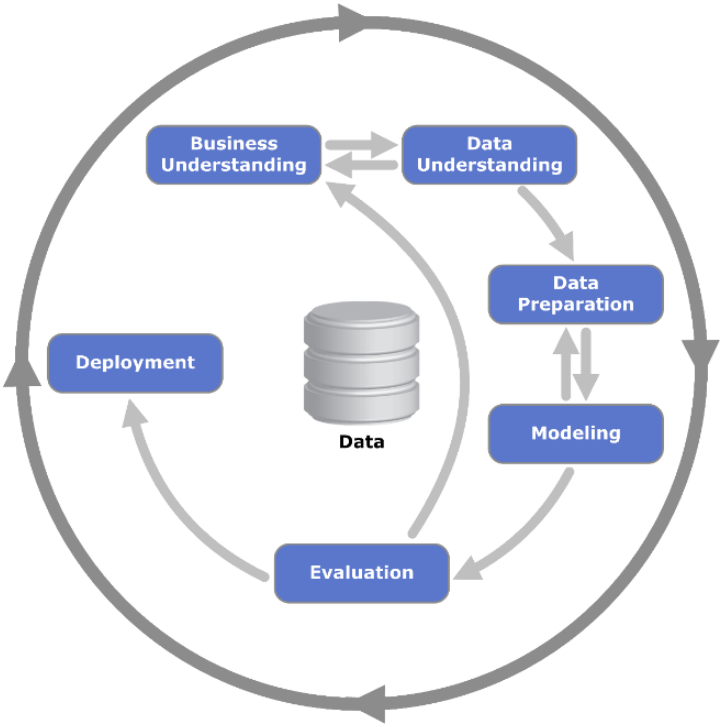


Figura 3.1: Modelo de referência de mineração de dados – CRISP-DM.

Entendimento do Negócio: essa fase inicial foca em entender o objetivo do projeto a partir de uma perspectiva de negócios, definindo um plano preliminar para atingir os objetivos.

Entendimento dos Dados: começa com a coleta de dados e início de atividades para familiarização com os dados, identificando problemas ou conjuntos interessantes para formular hipóteses sobre informações ocultas.

Preparação dos Dados: abrange as atividades para construção do conjunto de dados final. Normalmente ocorre várias vezes no processo, inclui seleção de tabelas, registro e atributo, bem como transformação e limpeza de dados para ferramenta de modelagem.

Modelagem: várias técnicas de modelagem são aplicadas, e seus parâmetros calibrados para otimização. Inúmeras técnicas podem ser empregadas para o mesmo problema de mineração de dados. Assim, é comum retornar à Preparação dos Dados durante essa fase.

Avaliação: é construído um modelo que parece ter alta qualidade sob uma perspectiva de análise de dados. No entanto, é necessário verificar se o modelo atinge os objetivos do negócio. Deve-se certificar se algum requisito de negócio deixou de estar presente no modelo.

Implantação: o conhecimento adquirido pelo modelo deve ser organizado e apresentado de uma maneira que o usuário possa empregar para suas necessidades de negócio.

Após efetivado todo o processo, busca-se obter um modelo que possa representar a realidade com maior exatidão.

Sob a ótica do arcabouço de gestão de riscos, definido na ISO 31000:2018 [15], cabe considerar que as mudanças de ambiente, seja por alteração de legislação, ou na forma de atuação das pessoas quando submetidas a fiscalização, ou até mudanças de panorama econômico, afetam o mundo real. Essas mutações refletem em inadequação do modelo gerado por não representar essa nova realidade.

Nesse contexto, o modelo deve ser reavaliado constantemente, assim como estabelece o método CRISP-DM, em razão da natureza cíclica da mineração de dados, e a norma ISO 31000, na aplicação dos princípios da "dinâmica" e da "melhoria contínua" ao processo de gerenciamento de riscos.

A atividade de "Tratamento de Riscos" da ISO 31000 foge do escopo desse trabalho, mas pode ser efetivada por medidas de decisão do Tribunal de Contas que interfere e altera nos controles existentes das atividades de servidores públicos, em casos pontuais ou até por norma de cumprimento obrigatória para todo complexo administrativo do DF.

3.2. Levantamento bibliográfico e resultados do enfoque meta-analítico de corrupção

A presente pesquisa utilizou a pesquisa bibliográfica de caráter exploratório e descritivo por meio do enfoque meta-analítico, elaborado por Mariano e Rocha [29], com adaptações.

A pesquisa da base bibliográfica utilizada nesta investigação considerou a busca pelo termo "*corruption civil servant*" nos documentos disponibilizados em diferentes fontes especializadas tais quais, *Web of Science*, retornando 149 artigos, *Scopus, com* 137 e outras bases científicas (*IEEE Xplore Digital Library* e *ACM Digital Library*).

Uma verificação simples da adequação dos artigos retornados da consulta à base de artigos da *Web of Science* é a nuvem de palavras das palavras-chave de todos os artigos coletados, disponível a seguir.

Figura 3.2: Nuvem de palavras dos artigos (Corrupção). Extraído do software *VosViewer 1.6.10*.

O resultado da frequência de palavras mostra-se adequado ao tema de pesquisa, ao identificarmos os termos *corruption, civil servants, governance, management,* entre outros, mas somente o exame dos documentos pode mostrar a adequação da investigação ao resultado esperado.

A literatura sobre o tema apresenta pressupostos e condições que favorecem a corrupção e sua influência na sociedade, mas raramente evidencia as características do comportamento dos servidores públicos em atos de corrupção.

Nesta pesquisa, essas características foram utilizadas, junto com a opinião de especialistas no assunto, para formação das trilhas que comporão a identificação de riscos de corruptibilidade dos servidores e aplicados como atributos para análise de mineração.

A pesquisa com enfoque meta-analítico apresentou os seguintes estudos relacionados a corrupção, descritos em ordem decrescente de citações segundo a *Web of Science* descritos na tabela a seguir.

Tabela 3.1: Publicações relacionadas a corrupção

Autores	Título / quant. citações	Contribuições
Anderson, CJ; Tverdova, YV [30]	Corruption, political allegiances, and attitudes toward government in contemporary democracies / (417)	A análise demonstra que os cidadãos em países com níveis mais altos de corrupção expressam avaliações mais negativas do desempenho do sistema político e exibem níveis mais baixos de confiança nos funcionários públicos. O efeito negativo da corrupção nas avaliações do sistema político é significativamente atenuado entre os que apoiam as autoridades políticas em exercício.
Liu, Bang-Cheng; Tang, Thomas Li-Ping [31]	Does the Love of Money Moderate the Relationship between Public Service Motivation and Job Satisfaction? The Case of Chinese Professionals in the Public Sector/(63)	Esse estudo identificou que existe moderação do gostar por dinheiro entre a motivação e a satisfação do trabalho, ou seja, o dinheiro é capaz de influenciar positivamente na satisfação e motivação no trabalho. Outro resultado foi que o gostar por dinheiro está relacionado a corrupção e comportamento antiético.
Burns, John P.; Wang Xiaoqi [32]	Civil Service Reform in China: Impacts on Civil Servants' Behaviour/(44)	Trata de reforma do serviço público na China que introduzindo processos de seleção mais competitivos, incentivos para recompensar o desempenho e reforçando o monitoramento e a supervisão, mas que em confronto com outras políticas que estavam sendo implementadas na época e por uma falha em abordar elementos da cultura organizacional recompensaram comportamentos ilegais, como a corrupção.
Hanna et al [33]	Dishonesty and Selection into Public Service: Evidence from India / (26)	A pesquisa versa sobre desonestidade e privilégios e indica que servidores que burlam em provas, "colam em provas", apresentam predição para comportamentos fraudulentos na administração pública daquele contexto.

Lassou et al. [34]	Government accounting reform in an ex-French African colony: The political economy of neocolonialism / (22)	Nesse estudo realizado em uma ex-colônia francesa os autores apontam como fatores de corrupção problemas relacionados à transparência, responsabilização e contabilidade governamental.
Asoni & Andrea [35]	Protection of property rights and growth as political equilibria / (22)	Aborda o tema de direitos de propriedade e crescimento econômico e como a falta de proteção desses direitos pode resultar em crescimento econômico lento devido à corrupção de funcionários públicos e outros fatores.
Dodge [36]	State and society in Iraq ten years after regime change: the rise of a new authoritarianism /(21)	O autor indica a corrupção política como fator que contribuiu para manutenção de um regime autoritário no Iraque.
Poocharoenet al. [37]	Meritocracy in Asia Pacific: Status, Issues, and Challenges /(18)	Trata de análise comparativa de sistemas de mérito entre os Estados Unidos, China, Coréia do Sul, Índia, Taiwan, Malásia e Filipinas com investigação em cinco dimensões: critérios de recrutamento; corrupção no recrutamento e promoção; filiação política e influência; nível de centralização dos processos de recrutamento e promoção; e a extensão dos regimes de proteção ao mérito.
Genaux, M [38]	Social sciences and the evolving concept of corruption / (17)	Trata da evolução do conceito de corrupção, desde os tempos de Aristóteles, do Império Romano, até os tempos modernos.
Rubbers, Benjamin [39]	The "informal sector": The economy of Katanga (Congo-Zaire) and the falsification of the law / (17)	Esse estudo aponta que, na República Democrática do Congo, a maioria das chamadas atividades "informais" envolvem a colaboração ativa de agentes do Estado, mobilizam uma lógica de corrupção nas redes sociais que ignoram as fronteiras institucionais. A corrupção, na maioria dos casos, se trata de um *continuum* entre favores recíprocos, quando eles pertencem à mesma rede de sociabilidade e extorsão, quando a distância social entre eles é máxima.
Liou, Kuotsai Tom; Xue, Lan; Dong, Keyong [40]	China's Administration and Civil Service Reform: An Introduction /(13)	Apresenta um panorama do desenvolvimento de políticas de reforma do funcionalismo público da China e examina desafios na avaliação do desempenho dos servidores, experiências de treinamento em serviço civil e reforma salarial no que se refere à corrupção pública.
Gong, Ting Ren, Jianming [41]	Hard Rules and Soft Constraints: regulating conflict of interest in China /(12)	Trata da definição de corrupção, de práticas de prevenção à corrupção na China, na relação entre conflitos de interesse e corrupção e discute sobre regulamentos e o comportamento eticamente sólido.

Fonte: *Web of Science* (atualizado em 17-ago-2020)

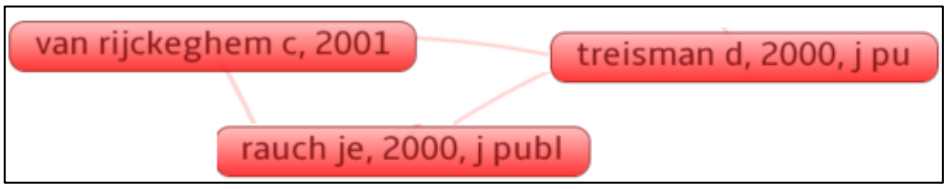
Percebe-se que os trabalhos mais citados tratam de corrupção sob diferentes perspectivas: quanto à nação como um todo ou entre nações, comportamento individual, cultura organizacional, relação de formalidade ou informalidade, quanto aos aspectos éticos e regulamentares.

Mapas de *co-citation* e *coupling*

Por meio do software VOSviewer 1.6.10 foram elaborados gráficos, facilitando a visualização da análise de *co-citation* e *coupling* relativa aos registros obtidos no *Web of Science* sobre corrupção de servidores públicos.

O exame de *co-citation* permite evidenciar os autores que são citados simultaneamente entre os artigos pesquisados. A análise de *coupling* identifica casos em que dois ou mais trabalhos referenciam outro trabalho em comum, restringindo-se aos últimos anos de publicação, temos como resultado quais as frentes de pesquisa que ainda continuam e predominam.

Apesar de não haver expressividade de artigos publicados quanto ao tema, o mapa de *co-citation* da figura 3.3 permite identificar três vertentes de estudo na área, representadas a seguir.



Fonte: autor. Extraído do software *VosViewer 1.6.10*.
Figura 3.3 – Visualização de rede de *co-citation*

Da pesquisa dos trabalhos representados no gráfico, pode-se identificar os autores com temas de estudo próximos, bem como os trabalhos relacionados com o tema proposto nesta dissertação.

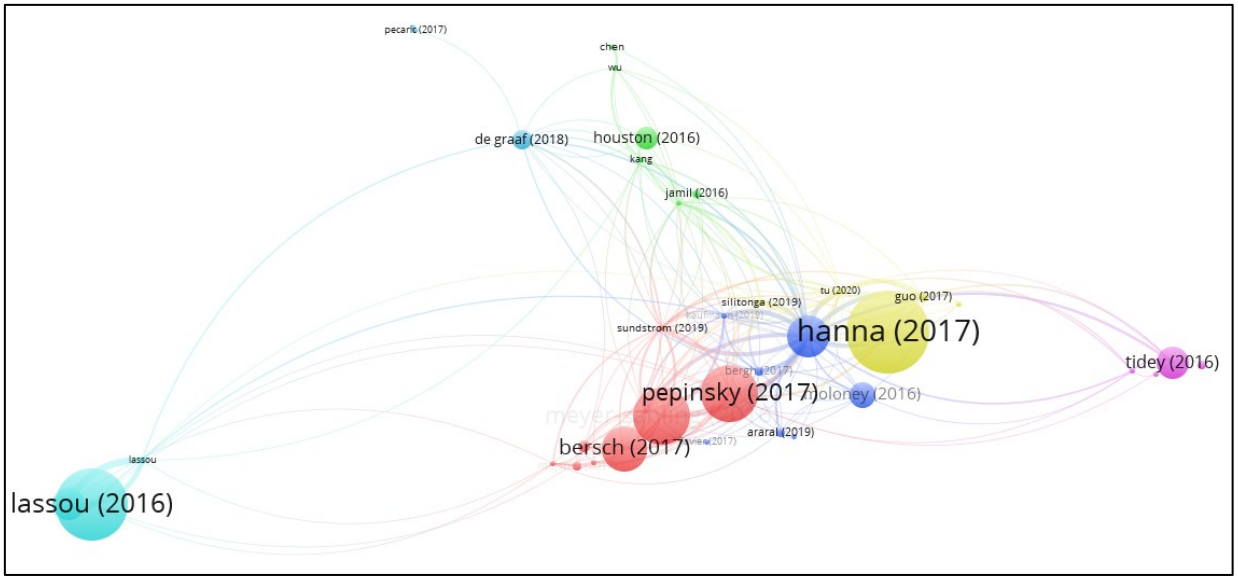
Em 2000, Treisman [42] apresentou um estudo amplo que trata de diversos aspectos de um país como, desenvolvimento da economia, estabilidade política, intervenção do Estado na economia, eficácia do sistema legal e salários, e concluiu como resultado de análise estatística que países predominantemente protestantes apresentam menor corrupção que os demais, do mesmo modo os países de democracia estabelecida há décadas, os países de sistema legal de *common law* e os estados unitários.

O mesmo autor, em estudo apresentado em 2007 [42], esclarece que há evidências que sugerem que democracias liberais altamente desenvolvidas e estabelecidas há muito tempo, com uma imprensa livre, com uma alta parcela de mulheres no governo e uma história de abertura ao comércio, são identificadas com menor corrupção. O autor acrescenta que países que dependem de exportações de combustíveis ou têm regulamentações comerciais intrusivas e inflação imprevisível são julgados mais corruptos.

Rijckeghem & Weder [43] tratam de corrupção sob o ponto de vista dos salários dos servidores, para isso utilizaram os salários de países de baixa renda. Como resultado identificaram relação estatística que indica menor corrupção para salários mais altos.

Rauch & Evans [44] investigam características institucionais de burocracia e níveis de corrupção e conclui que os fatores salários competitivos, promoção interna, estabilidade de carreira e recrutamento meritocrático são relevantes para melhorar a burocracia e permitir o crescimento econômico do país.

Conforme ensina Mariano e Rocha [29], a análise de *coupling* (figura 3.4) mostra as principais frentes de pesquisa.



Fonte: autor. Extraído do software *VosViewer 1.6.10*.
Figura 3.4 – Representação de *Coupling*.

A análise de *coupling* identifica as abordagens mais recentes. E segundo o gráfico, Hanna [33], Lassou[34], Pepinsky[45] e Bersch & Matthew[46] estão em maior evidência.

No estudo de Hanna [33] avalia-se a desonestidade e privilégios no serviço público da Índia e apontou-se como resultado que os servidores indianos que burlam em provas apresentam predição para comportamentos fraudulentos na administração pública daquele país. Outro achado importante é que a variação de corrupção, para indivíduos sob mesmos contextos e incentivos, pode ser alavancada pela propensão a desonestidade do indivíduo. E por fim sugere a melhoria do processo de seleção de servidores públicos, não somente sobre o aspecto da capacidade, para melhor excluir indivíduos com propensão a corrupção.

Pepinsky et al.[45] apresentam um estudo sobre burocracia e prestação de serviço nos países em desenvolvimento e concluem que as burocracias voltadas a concessão de licenças e

permissões devem gerar mais oportunidades de suborno e as burocracias de criação criam oportunidades de corrupção durante as licitações e contratações.

Outro estudo realizado, em uma ex-colônia francesa, Lassou[34] referencia que a fraca transparência, responsabilização e contabilidade governamental contribuem para corrupção, fraca governança e baixo desenvolvimento de um país.

Bersch & Matthew [46] tratam da capacidade do Estado, da politização da burocracia e corrupção no Brasil. Nesse estudo os autores utilizaram como base mais de 260 mil servidores civis de agências federais e concluem, entre outros fatores, que o domínio partidário de nomeações políticas está associado a menor capacidade de agência, que é importante para o funcionamento cotidiano da burocracia e para a capacidade da burocracia de combater efetivamente a corrupção.

Outra pesquisa relevante de *coupling* é de Meyer-Sahling & Mikkelsen [47], a qual relaciona direito, mérito, política e corrupção sob a perspectiva dos funcionários públicos de cinco países do leste europeu. Nesta pesquisa identificou-se que quanto maior a politização, indicação política de cargos, maior é o nível de corrupção; enquanto, quanto maior o mérito do recrutamento, menor é a corrupção associada.

Para o presente trabalho de pesquisa, utilizou-se a segmentação em quatro dimensões, assim como apresentado no trabalho de Carvalho [48], conforme a seguir.

Tabela 3.2 – Principais fontes por dimensão de pesquisa para estabelecer possíveis atributos para mineração de dados em corrupção de servidores públicos.

DIMENSÃO CORRUPÇÃO	
Servidores demitidos com base na Lei nº 8.429/92 e com penalidades no TCDF	Hanna [33], Carvalho & Carvalho [49], Carvalho [48],
DIMENSÃO FUNCIONAL	
Renda	Gans-Morse [10], Kuotsai [40], Carvalho [48]
Patrimônio	Gans-Morse [10]
Cargo público (atribuições)	Padula & Albuquerque [7], Poocharoen [37], Carvalho & Carvalho [49] Carvalho [48]
Servidor sócio de empresas com contrato com o GDF	Carvalho [48], sugestão de especialistas
Servidor beneficiário de programas de governo	Valcárcel [50]
DIMENSÃO POLÍTICA	
Servidores com relações partidárias	Pedersen [51], Bersch & Matthew[46], Meyer-Sahling & Mikkelsen [47], Moro [6], Carvalho & outros [52], Carvalho [48]
Transparência do órgão do servidor	Lassou[34], Treisman [42], Gans-Morse [10]
DIMENSÃO SOCIETÁRIA	

Servidor sócio de empresas contratadas pelo GDF	Carvalho [48], sugestão de especialistas.

Essas dimensões de atuação foram foco de análise a partir das diversas bases de dados governamentais do DF e de outros entes da federação disponíveis para uso do TCDF e foram utilizados para formação inicial dos atributos para o processo de mineração de dados.

3.3. Levantamento bibliográfico e resultados do enfoque meta-analítico para técnicas de mineração

Para identificação das abordagens de mineração de dados para o tema de pesquisa, utilizou-se, assim como no levantamento anterior, a pesquisa bibliográfica de caráter exploratório e descritivo por meio do enfoque meta-analítico, elaborado por Mariano & Rocha [29], com adaptações.

A pesquisa da base bibliográfica utilizada nesse item considerou a busca pelo termo "*data mining corruption civil servant*" nas fontes especializadas *Web of Science* e *Scopus* resultando em apenas um artigo: *Using political party affiliation data to measure civil servants' risk of corruption* [52].

De forma a ampliar a pesquisa, reduziu-se os termos para "*data mining corruption*" que retornou 78 artigos da *Web of Science* e 130 artigos da *Scopus* restringindo-se aos últimos dez anos.

Apesar do aumento do quantitativo de documentos obtidos, cabe frisar que a maioria desses artigos estão relacionados a corrupção de imagens, memória ou de dados e ainda, de empresas de mineração de pedras, restando poucos artigos relativos à corrupção de pessoas, que permitam compreender o contexto da mineração de dados para o tema objeto dessa dissertação.

A revisão de literatura identificou a aplicação de mineração de dados em diversos setores para combate ou análise de corrupção e fraudes.

Na área de saúde identificou-se o uso de redes neurais [53] para avaliar a percepção de corrupção entre países. Um estudo de revisão de literatura [54], indicou diversos artigos técnicos que usam técnicas de mineração de dados para combate à fraude no setor de saúde.

Na área financeira, registram-se diversos algoritmos de mineração para diferentes aplicações: na detecção de lavagem de dinheiro utilizou-se de *clustering* e *k-means* [55] e redes

neurais [56]; para detecção de fraude financeira empregou-se redes neurais e algoritmo de classificação *support vector machine* – SVM [57].

No Brasil, um estudo para identificação de cartéis em licitação [58] empregou *clustering* e regras de associação; para avaliação de risco de corrupção utilizou-se *Naive Bayes* [59]; para identificação de anomalias em compras governamentais de TI fez-se uso de *deep learning* [60]; para avaliação de dependência de filiação partidária com corrupção utilizou-se redes neurais com *backpropagation*, redes Bayesianas, *Support Vector Machines* e *Random Forest* [52] e outro estudo recorreu às redes Bayesianas para avaliação de risco de corrupção nas unidades de gerenciamento governamental [49].

Uma pesquisa na Macedônia [61] identificou regras de associação e *clustering* para identificação de desvios nas compras governamentais. Na China [62], um sistema de prevenção para risco de corrupção foi desenvolvido com uso de redes neurais, *clustering* e análise de desvios ou anomalias, os mesmos autores também elaboraram um trabalho científico [63] de modelo de sistema de risco de corrupção baseado em nuvem.

Na Indonésia aplicou-se *Naive Bayes* [64] para detecção de fraude no sistema de compras governamentais e em Bangladesh utilizou-se *k-means* modificado para elaboração de modelo de detecção de corrupção [65].

Também se identificou estudos que apresentam exames de corrupção por mineração na web [66] [67]. No entanto, na presente investigação restringiu-se aos dados abertos, sem necessidade de mineração na Web, e aos dados de caráter sigilosos presentes em bancos de dados governamentais distritais e federais.

O quadro a seguir apresenta uma síntese desses trabalhos por técnica de mineração aplicada.

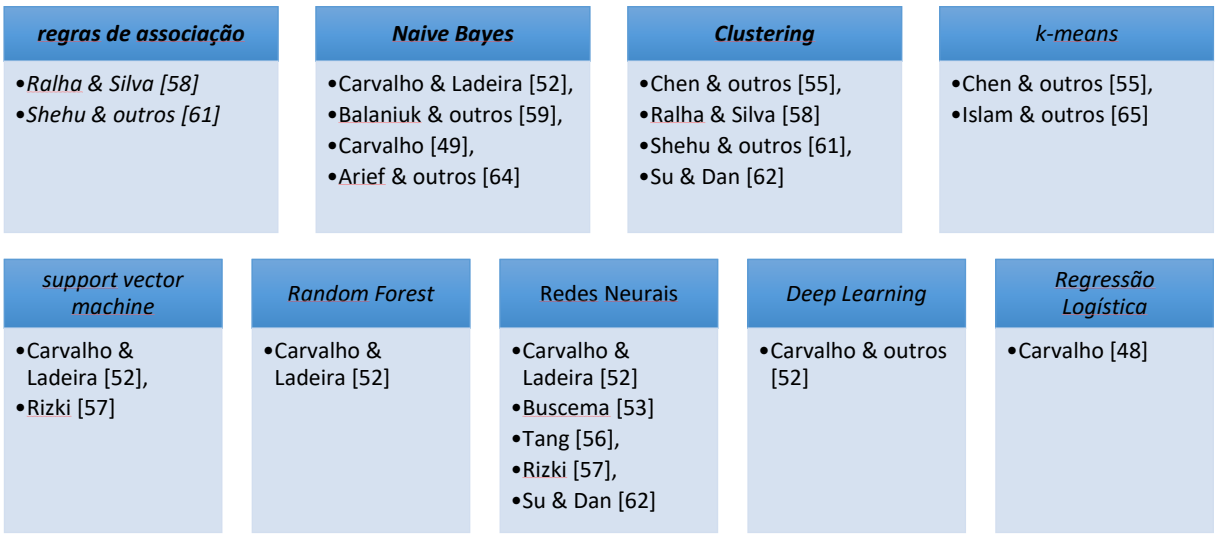


Figura 3.5 – Técnicas de aprendizagem de máquina e estudos relacionados a corrupção.

A depender do problema de aprendizado de máquina a ser solucionado podem ser empregadas diferentes técnicas conforme se denota no quadro anterior. Essas técnicas são categorizadas como supervisionadas ou não supervisionadas.

Na aprendizagem supervisionada, o objetivo é prever o valor de uma medida de resultado com base em várias medidas de entrada; na aprendizagem não supervisionada, não há medida de desfecho, e o objetivo é descrever as associações e padrões entre um conjunto de medidas de entrada [68].

Do quadro, temos como técnicas supervisionadas *Naive Bayes*, regressão, redes neurais, *support vector machine* e *random forest* e como técnicas não supervisionadas: regras de associação, *clustering* ou agrupamento e *k-means*.

Como trabalhos mais afetos à presente pesquisa consideram-se "*Using Political Party Affiliation Data to Measure Civil Servants' Risk of Corruption*" [52] e "*Bayesian Models to Assess Risk of Corruption of Federal Management Units* [49]" e a dissertação de mestrado apresentada neste PPCA de título "Modelos Preditivos para Avaliação de Risco de Corrupção de Servidores Públicos Federais [48]".

Ao analisar os documentos relevantes com *co-citation* das bases *Web of Science* e *Scopus* para os termos pesquisados, não obtivemos como resultado autor que trate de mineração de dados relacionada à corrupção de pessoas.

Da mesma forma ocorreu para análise de *Coupling* que poderia evidenciar os possíveis *fronts* de pesquisa, mas devido à variabilidade semântica dos termos empregados, resultou em

casos que tratam de empresas de mineração de pedras e, também, de corrupção de imagens, de memória ou de dados.

Observam-se diversos métodos para avaliação de fraude e corrupção, para o presente trabalho, dada a natureza de a variável dependente ser dicotômica, ou seja, binária, optou-se pelo uso de regressão logística, que será abordada no item 2.2.4.

3.4. Balanceamento de dados

Em diversas situações de mineração de dados ocorre uma característica de desbalanceamento de classes, isso pode ser definido como um cenário em que os dados são preponderantemente de uma classe, enquanto a classe considerada relevante para o estudo apresenta poucos casos e é denominada **classe rara** [69].

Um exemplo de classe rara pode ser em um contexto em que a quantidade de transações fraudulentas de cartão de crédito atinge 1% em relação ao total de transações. A classe que deve ser examinada, de transações fraudulentas, representa um reduzido quantitativo de casos.

Para casos em que a classe rara é a mais importante, o custo de associação errônea na classe rara é maior que da associação inverídica na classe comum. Para exemplificar, podemos citar o exemplo anterior: considerar uma atividade fraudulenta de cartão de crédito de forma incorreta irá resultar na negativa da transação comercial, mas uma atividade efetivamente fraudulenta tida honesta permitirá a conclusão da fraude.

Se o objetivo é avaliar a classe rara, então é necessário estabelecer alguma forma de tratar os dados para obter a melhor performance na execução dos algoritmos de aprendizado de máquina e obter um modelo mais adequado a realidade, pois os algoritmos tendem a apresentar mais erros de classificação na classe rara do que na classe majoritária, quando os dados estão desbalanceados [70].

Duas abordagens possíveis para aplicação de aprendizagem de máquina em dados desbalanceados são [70]:

- ✓ solução com algoritmo otimizado segundo características específicas para cenários de desbalanceamento de classes ou;
- ✓ solução com tratamento de dados, balanceamento dos dados com técnicas de reamostragem para adequado treinamento dos algoritmos.

O emprego de algoritmos de aprendizagem de máquina no contexto de desbalanceamento expressivo de classes ocorre em condições em que à prevalência da classe minoritária, que é a classe de interesse, devem ser atribuídas condições de possibilidade de maximizar os resultados, para isso, a atribuição de pesos é medida eficaz nesse contexto, outros requisitos podem ser empregados como forçar a estratificação equilibrada da classe de interesse nas bases de treino e teste além de outras medidas aplicadas no âmbito dessa pesquisa [71].

Quanto ao pré-processamento, dois métodos são considerados usuais para minimizar o desbalanceamento de classes na fase de pré-processamento de dados: *undersampling* e *over-sampling*. O primeiro trata da exclusão randômica de observações da classe majoritária, enquanto o segundo da criação múltipla de cópias de observações da classe minoritária, mas os dois métodos apresentam desvantagens, o *undersampling* pode descartar instâncias de dados potencialmente úteis, enquanto o *over-sampling* pode aumentar a probabilidade de *overfitting,* que corresponde a ocorrência de um modelo estatístico muito bem ajustado ao conjunto de dados anteriormente observado, mas se mostra ineficiente para prever novos resultados.

Para minimizar os efeitos das técnicas anteriores, foi criada a técnica *SMOTE, synthetic minority oversampling technique*[23]*,* que seleciona aleatoriamente um ou mais vizinhos mais próximos de uma instância de classe minoritária e produz novas instâncias com base nas interpolações lineares entre as instâncias originais e os vizinhos mais próximos selecionados aleatoriamente [70].

Segundo os autores, a técnica SMOTE foi elaborada para contexto de fraude, em que se caracteriza por uma classe relevante minoritária, e para que possa ser aplicado em técnicas de aprendizagem de máquina deve-se proceder ao equilíbrio de classes para evitar a ocorrência de viés de predição para classe dominante nos algoritmos empregados. Ao equilibrar as classes por SMOTE os algoritmos podem ser treinados e quando aplicados aos dados para exame de predição os resultados não serão tendenciosos para a classe dominante.

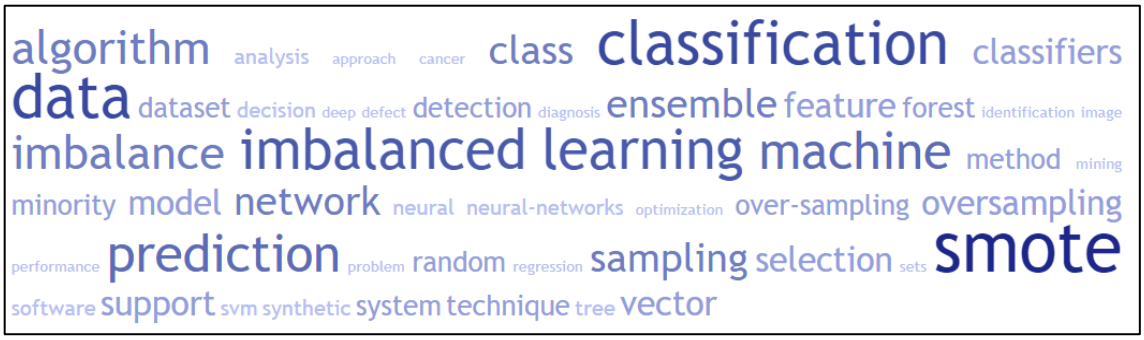
A técnica procede a combinação de super-amostragem da classe minoritária, sub-amostragem da classe majoritária e criação de exemplos sintéticos de classe minoritária.

Para identificação da base bibliográfica relativas à técnica SMOTE buscou-se o termo “SMOTE” na *Web of Science*, retornando-se 946 artigos para os últimos cinco anos (2016 a 2020).

Fonte: Extraído do *Web of Science.*
Figura 3.6 – Gráfico de publicações por ano.

Na figura 3.6, apresenta-se a quantidade de itens publicados ano a ano. Pode-se perceber o aumento gradativo da quantidade de artigos publicados até maio de 2020. Segundo Mariano e Rocha [72], esta análise é importante e está baseada na Teoria Epidêmica de Goffman, que afere a razão de crescimento e declínio de determinada área de conhecimento.

Uma verificação simples da adequação dos artigos retornados da consulta a base de artigos da *Web of Science* é a nuvem de palavras das palavras-chaves de todos os artigos selecionados.



Fonte: autor. Extraído do software *VosViewer 1.6.10*.
Figura 3.7 – Nuvem de palavras dos artigos (SMOTE).

O resultado da frequência de palavras mostra-se adequado às aplicações da técnica SMOTE, por evidenciar palavras do contexto próprio da presente pesquisa, mas somente o exame dos documentos obtidos pode mostrar a adequação da investigação ao resultado esperado.

A presente investigação utilizou a pesquisa bibliográfica de caráter exploratório e descritivo por Mariano & Rocha [29], com adaptações.

Segundo a abordagem do enfoque meta-analítico [29], cabe descrever os estudos relacionados a técnica SMOTE, que estão descritos em ordem decrescente de citações conforme tabela a seguir.

Tabela 3.3 – Publicações relacionadas a técnica SMOTE

Autores	**Título / Quantidade de Citações**	**Contribuições**
Branco & al. [73]	A Survey of Predictive Modeling on Imbalanced Domains / 113	Apresenta um levantamento das técnicas existentes para lidar com as aplicações da análise preditiva no contexto de desbalanceamento, descreve as principais abordagens e propõe uma taxonomia dos métodos.
Buda et al. [74]	A systematic study of the class imbalance problem in convolutional neural networks / 112	Apresentam uma investigação sistemática do impacto do desequilíbrio de classe no desempenho de classificação de redes neurais convolucionais (CNNs) e uma comparação dos métodos frequentemente usados para abordar a questão.
Fernández et al. [75]	SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary / 61	Neste artigo os autores apresentam o contexto da técnica SMOTE após 15 anos de sua criação, contexto atual e quais os desafios existentes e as implementações em pacotes para softwares.
Mao et al. [76]	Online sequential prediction of bearings imbalanced fault diagnosis by extreme learning machine / 61	O artigo trata do problema de métodos tradicionais apresentarem baixa precisão no diagnóstico de falhas em rolamentos e os autores propõem uma solução por processo de aprendizagem de máquina online.
Sun et al. [77]	Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates / 55	Trata-se de uma aplicação em um banco na China para realizar avaliação de crédito com uso de árvore de decisão com dados desbalanceados.
Sáez et al. [78]	Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets / 54	O artigo trata de um estudo sobre superamostragem para conjuntos de dados desequilibrados de várias classes. Esclarece que para duas classes existem várias abordagens, mas para mais de duas a solução é complexa.

Percebe-se que os trabalhos mais citados tratam da técnica SMOTE em diversas situações e perspectivas.

Em [73] os autores propõem uma nova taxonomia para as abordagens existentes, agrupando-as em (i) pré-processamento de dados, (ii) métodos de aprendizado para fins especiais, (iii) pós-processamento de previsão e (iv) estratégias híbridas.

Buda et al. [74] informam que para aprendizagem de máquina muitas investigações foram realizadas, mas para aprendizado profundo (*deep learning*) ainda é muito limitado. Com o estudo de três conjunto de dados de referência de crescente complexidade para investigar os efeitos de desequilíbrio na classificação e concluíram que (i) o efeito do desequilíbrio de classe no desempenho da classificação é prejudicial; (ii) o método de abordagem do desequilíbrio de classe que emergiu como dominante em quase todos os cenários analisados foi a superamostragem; (iii) a sobreamostragem deve ser aplicada ao nível que elimina completamente o desequilíbrio, enquanto a taxa ideal de subamostragem depende da extensão do desequilíbrio; (iv) ao contrário de alguns modelos clássicos de aprendizagem automática, a superamostragem não causa o super ajuste das CNNs; (v) o limiar deve ser aplicado para compensar as probabilidades da classe anterior quando o número geral de casos classificados adequadamente for de interesse.

Em [75] os autores apresentam o cenário após quinze anos da criação da técnica SMOTE, uma análise das variações da técnica criadas, as melhorias nas diferentes formas e desvantagens detectadas em relação ao modelo original de SMOTE.

Também relatam a aplicação potencial a problemas de previsão mais complexos, como dados de streaming, aprendizado semi-supervisionado, múltiplas instâncias e regressão.

Os autores aditam os desafios atuais, com destaque à necessidade de aprimorar o tratamento de ruído, falta de dados, sobreposição, mudança de conjunto de dados e a maldição da dimensionalidade e sugerem como relevante o foco em desbalanceamento nas estruturas de Big Data e processamento em tempo real.

Em [76], os autores esclarecem que em muitas aplicações reais de diagnóstico de falhas de rolamentos, os dados *on-line* tendem a ser desequilibrados, o que significa que o número de dados de falhas é muito menor que os dados normais enquanto todos são coletados de maneira sequencial on-line.

Nesse cenário, os autores propuseram um método de previsão sequencial on-line para o problema de diagnóstico de falhas desequilibradas, com base na aprendizagem de máquina

extrema com o uso de SMOTE e os resultados demonstraram que o método proposto pode melhorar a precisão do diagnóstico de falhas com melhor eficácia e robustez.

Um novo modelo árvore de decisão efetiva para avaliação de crédito empresarial foi proposto por [77] com base na técnica de sobreamostragem minoritária sintética (SMOTE) e o algoritmo de aprendizado *Bagging ensemble learning algorithm with differentiated sampling rates* (DSR). Como resultado, os autores identificaram a eliminação do problema de desequilíbrio de classe da avaliação de crédito e melhora na precisão positiva.

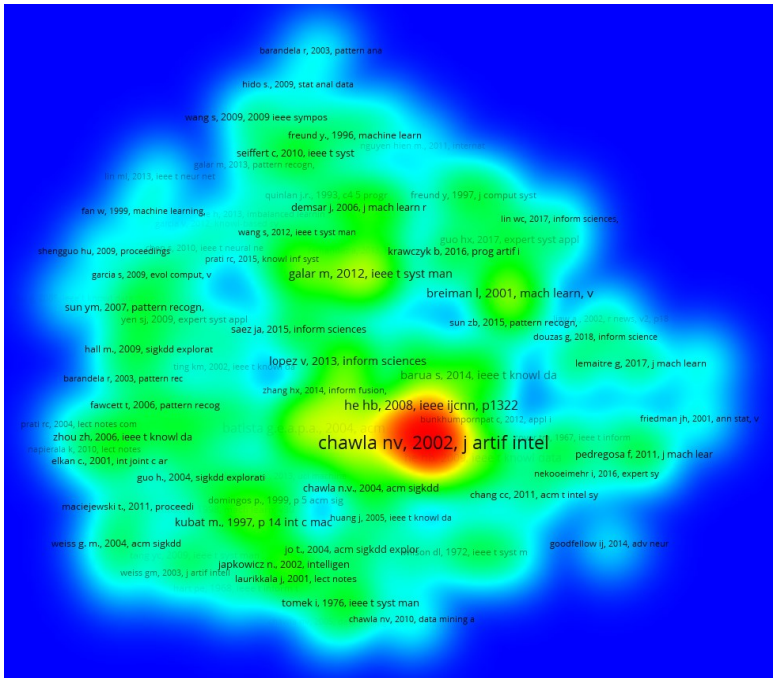
Em [78], os autores elaboraram um estudo sobre superamostragem para conjuntos de dados desequilibrados de várias classes. Relatam que, nos últimos anos, muitas soluções foram propostas para combater a classificação desequilibrada, mas elas se concentram principalmente em cenários binários.

O resultado dessa pesquisa foi obter uma visão mais profunda da natureza dos problemas de dados desequilibrados de várias classes e apresentar tipos de exemplos presentes que podem ser usadas no futuro para projetar novos algoritmos de aprendizado de pré-processamento que incorporarão esse conhecimento básico sobre a estrutura do problema.

Por meio do software VOSviewer 1.6.10 foram elaborados mapas de calor e rede de relacionamento, facilitando a visualização da análise de *co-citation* e *coupling* relativa aos registros obtidos no *Web of Science* sobre a técnica SMOTE.

O exame de *co-citation* permite evidenciar os autores que são citados simultaneamente entre os artigos pesquisados. A análise de *coupling* identifica casos em que dois ou mais trabalhos referenciam outro trabalho em comum, restringindo-se aos últimos anos de publicação, temos como resultado quais as frentes de pesquisa que ainda continuam e predominam.

O mapa de *co-citation* (figura 3.8) permite identificar uma vertente de estudo na área, representadas em cores diferentes.

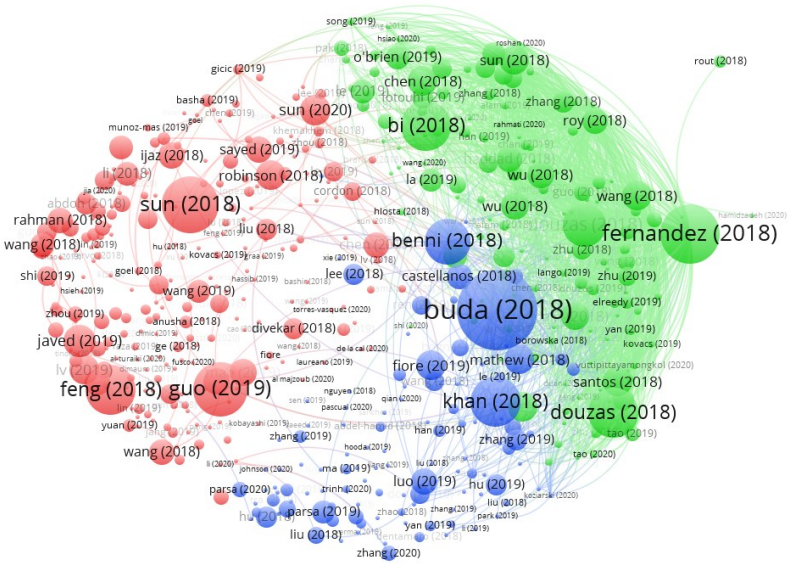


Fonte: autor. Extraído do software *VosViewer 1.6.10*.
Figura 3.8 – Mapa de calor de *co-citation (SMOTE)*.

Conforme o mapa de co-citation (figura 3.8), existe uma convergência das citações para o artigo de Chawla "*SMOTE: Synthetic Minority Over-sampling Technique*"[23] que divulgou a origem da técnica.

Segundo os autores, o artigo busca mostrar que a combinação do método de superamostragem da classe minoritária (anormal), subamostragem da classe majoritária (normal) com criação de exemplos sintéticos de classe minoritária pode obter melhor desempenho do classificador (no espaço ROC) do que apenas subamostragem da classe majoritária. A técnica está a ser empregada cada vez mais, até agosto de 2020 o artigo apresentou 5.086 citações.

Conforme ensina Mariano e Rocha [29], a análise de *coupling* mostra as principais frentes de pesquisa, conforme figura a seguir



Fonte: autor. Extraído do software *VosViewer 1.6.10*.
Figura 3.9 – Representação de *Coupling (SMOTE)*.

Segundo essa análise, as abordagens mais recentes estão separadas por cores na figura 3.9. Três frontes de investigação são delineadas e apresentam como principais artigos os seguintes: Fernandez [75]; Buda [74] outro cluster formado pelos artigos de Sun [77], Feng et al. [79] e Guo [80].

O artigo de Fernandes [75] foi comentado anteriormente entre os artigos com destaque em citações, este artigo apresentou a evolução da técnica SMOTE nos 15 anos de criação aos dias atuais.

A outra possível frente de investigação apresenta o artigo de Buda [74] como expoente nesse *cluster (de cor azul)*, também foi comentado entre os artigos com mais citações e tratou de uma investigação sistemática do impacto do desequilíbrio de classe no desempenho de classificação de redes neurais convolucionais (CNNs).

O último ‘cluster’ identificado apresentou três artigos em destaque: Sun [77] sendo comentado entre os artigos em destaque de citações e tratou de uma aplicação em um banco na China para avaliação de crédito com dados desbalanceados; Feng et al. [79] que tratou de aplicação em medicina na análise quantitativa com aprendizagem de máquina aplicada em imagens tomográficas de pequenas massas renais e Guo [80], que tratou com *extreme learning machine* (ELM) e CR-SMOTE, um algoritmo modificado do SMOTE, que estabelece melhoria no relatório de bugs de teste de software no contexto de *crowdsourced*.

Implementação da técnica SMOTE na linguagem R

Um grupo de pesquisa da Universidade do Porto disponibilizou um pacote para implementação da técnica SMOTE no software de estatística R com características que permitem diversas parametrizações.

Esta função está disponível no pacote DMwR [81] de funções para mineração de dados para o software R, sob autoria de Luís Torgo, da Universidade do Porto disponível no repositório CRAN e melhor exemplificado no livro do mesmo autor "*Data Mining with R - Learning with Case Studies*" [82].

SMOTE (form, data, perc.over = 200, k = 5, perc.under = 200, learner = NULL, ...)

Os argumentos disponíveis permitem estabelecer o grau percentual de reamostragem de *undersampling* pelo argumento "perc.under" e de *over-sampling* pelo argumento"perc.over". Outro parâmetro, "k" define a maneira como novos exemplos serão criados com base no algoritmo *k nearest neighbours* (KNN). O último parâmetro disponível é da função de aprendizado "learner", se definido como NULL apresentará o modelo de classificação presente na função SMOTE, caso contrário, o modelo que estiver definido pelo usuário.

Para a presente pesquisa a técnica *synthetic minority oversampling technique SMOTE* [23] mostra-se relevante por tratar de forma adequada o problema de classes desbalanceadas principalmente quando a classe de interesse representa um valor próximo ou menor que 1%.

3.5. Regressão Logística

A Regressão Logística - RL é um método elaborado sob a liderança do estatístico Sir. Ronald Fisher e envolve a estimativa de parâmetro β de uma distribuição de probabilidade de variável aleatória X com determinado número de observações independentes.

O emprego da RL é usual em situações em que a variável dependente é de natureza binária ou dicotômica, enquanto as variáveis independentes podem ser categóricas ou não. A RL busca estimar a probabilidade associada à ocorrência de determinado evento em relação a um conjunto de variáveis que explicam o fenômeno.

Na RL a probabilidade de ocorrência de um evento pode ser estimada diretamente. A variável dependente pode assumir dois estados (0 ou 1) e haver um conjunto de *p* variáveis independentes X1, X2, …, Xp conforme a equação a seguir.

$$P(Y = 1) = \frac{1}{1 + e^{-g(x)}}$$

Onde g(x) = Bo+B1X1+…+BpXp

Os coeficientes B0, B1, …Bp são estimados a partir do conjunto de dados, pelo método da máxima verossimilhança, que determina a combinação de coeficientes que maximiza a probabilidade.

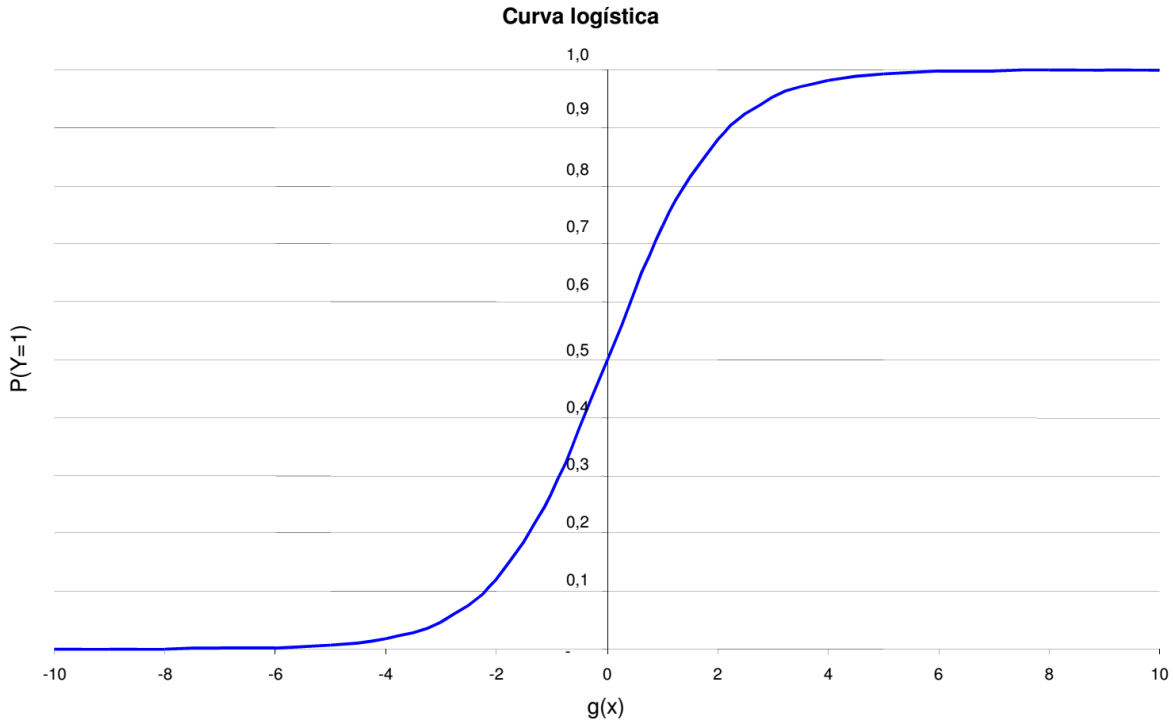


Figura 3.10 – Curva característica de uma regressão logística.

A interpretação dos coeficientes da RL para valores positivos representa aumento da probabilidade e negativo a redução da probabilidade.

As referências clássicas de regressão logística são Cox & Snell e Hosmer & Lemeshow [83].

3.6. Métricas de Validação

Para o processo de avaliação de um classificador, no caso em estudo considerar corrupto ou não, um valor binário, a forma mais compreensível é pela matriz de confusão.

Essa matriz, para o caso de classificação binária, apresenta dimensão 2 x 2 e exibe os resultados do modelo de predição ao mostrar o número de classificações corretas e das classificações preditas [84], conforme a descrição a seguir.

Tabela 2.4 – Matriz de Confusão para classificação binária

Classe esperada	**Classe predita**	
	Previsão Positiva	**Previsão Negativa**
Positiva	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Negativa	Falso Positivo (FP)	Verdadeiro Negativo (VN)

As células da matriz de confusão podem indicar problemas mais significativos ou não nos resultados do classificador, a depender do contexto em exame.

Para a presente pesquisa, o caso VP representa o "Corrupto" e VN o "não corrupto". Os demais casos Falso Positivo e Falso Negativo estão associados ao erro de predição que em conjunto com os demais valores permite fazer avaliações de métrica de validação do modelo a seguir descritas [84].

Acurácia – é a medida mais utilizada para avaliação de classificadores, também denominada taxa de classificações corretas, descrita pela fórmula a seguir.

$$Acurácia = \frac{(VP + VN)}{(VP + FP + VN + FN)} \tag{2.2.9.1}$$

Determina o número de predições corretas dividido pelo número de amostras totais (a soma das entradas da matriz de confusão).

Precisão ou preditiva positiva – é a porcentagem de acertos ou verdadeiros positivos (VP) em relação a todos os exemplos de classificados como positivo (VP + FP).

$$Precisão = \frac{(VP)}{(VP + FP)} \tag{2.2.9.2}$$

Especificidade ou taxa de verdadeiros negativos – é a taxa de rejeições corretas, isto é, a porcentagem de verdadeiros negativos (VN) em relação a todos os exemplos em que a classe esperada é a classe negativa (FP+VN).

$$Especificidade = \frac{(VN)}{(FP + VN)} \tag{2.2.9.3}$$

Sensibilidade (do inglês *recall*) – também conhecida por taxa de verdadeiros positivos, pois é a porcentagem de verdadeiros positivos (VP) em relação a todos os positivos da predita (VP+VN).

$$Sensibilidade = \frac{(VP)}{(VP + FN)} \tag{2.2.9.4}$$

F-score – Essa medida é calculada pela média harmônica das medidas de precisão e sensibilidade.

$$F1 - measure = \frac{2}{\left(\frac{1}{sensibilidade}\right) + \left(\frac{1}{precisão}\right)} \tag{2.2.9.5}$$

A equação anterior pode ser reescrita da seguinte forma:

$$F1 - measure = \frac{2 \cdot sensibilidade \cdot precisão}{(sensibilidade + precisão)} \tag{2.2.9.6}$$

F-measure é uma medida capaz de avaliar em conjunto a sensibilidade e precisão, mostra-se uma referência adequada para estudo de classes raras, pois o cenário de alta precisão e baixa sensibilidade pode ser avaliado por essa grandeza.

A medida de acurácia do classificador é normalmente considerada como o principal valor na métrica de validação. No entanto, como para o caso de classificador binário existe a possibilidade de apresentar alta acurácia e ainda assim os erros do modelo serem representativos, indica que o classificador não responde adequadamente no modelo. Por esse motivo, as outras medidas devem ser utilizadas para avaliar o modelo como um todo.

Sensibilidade (do inglês *recall*) – Mandrekar [24], em seu estudo, apresenta como avaliação de desempenho de preditor binário uma curva de característica de operação do receptor (*receiver operating characteristic* - ROC) que inclui todos os limites de decisão possíveis de um resultado de teste de diagnóstico.

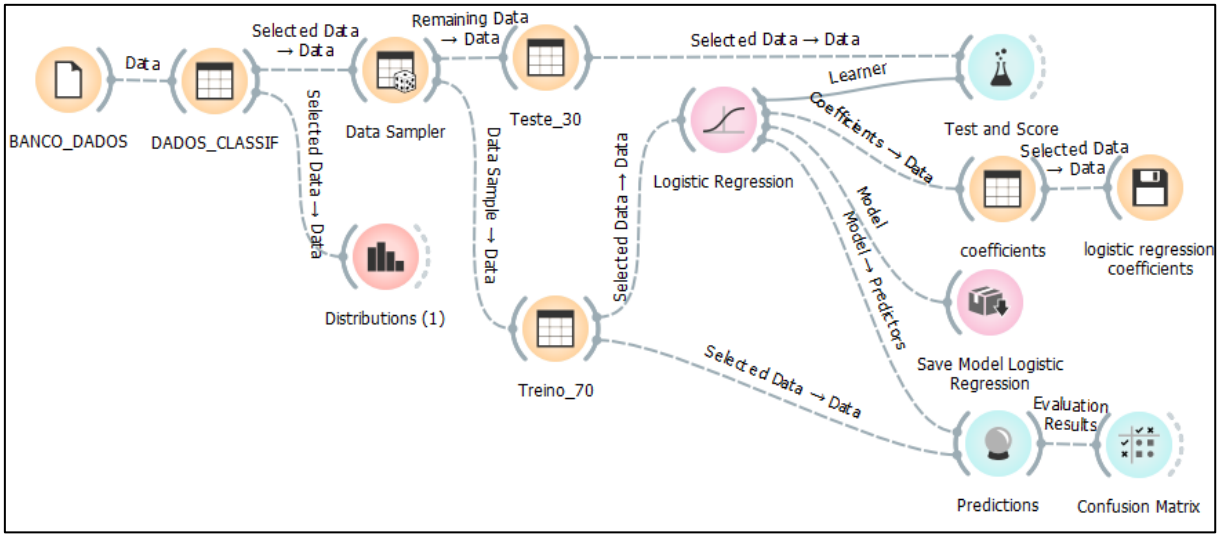
Mandrekar [24] definiu como curva ROC um gráfico de sensibilidade em relação à especificidade de um teste e a medida de área sobre essa curva expressa a medida de desempenho do modelo.

Nesse estudo estabelece-se que uma área sobre a curva ROC – AUC, (*Area Under Curve – AUC*) de 0,5 sugere nenhuma discriminação; 0,7 a 0,8 é considerado aceitável; 0,8 a 0,9 é considerado excelente e mais de 0,9 é considerado excepcional.

4. Solução proposta

O universo da pesquisa representa a totalidade de servidores do Governo do Distrito Federal que representa cerca de 261 mil servidores, inclusive aposentados, e como atributo alvo temos 294 servidores punidos por atos de improbidade nos termos da Lei nº 8.429, de 2 de julho de 1992 (cerca de 0,14% da população dos servidores do Governo do Distrito Federal).

A proposta de solução segue o método CRISP-DM e para elaboração do modelo orientou-se pelo diagrama a seguir.



Fonte: Elaborado pelo autor com uso do Orange Canvas
Figura 4.1 – Diagrama das etapas da solução proposta.

Em geral, o processo de aprendizagem de máquina ocorre com a construção de modelo e depois aplicação do modelo para avaliação de novos casos. Para presente pesquisa, busca-se apenas a identificação dos fatores de risco que corresponde a obtenção dos coeficientes da regressão logística que traduzem quais atributos (ou variáveis) são considerados relevantes do ponto de vista de gerenciamento de riscos.

Os especialistas que participaram da pesquisa são auditores com conhecimento de negócio e/ou dados que atuam na fiscalização em diversos setores: dois na área de legislação de pessoal; um auditor com experiência na fiscalização de prestação de contas partidárias junto ao Tribunal Superior Eleitoral; três auditores do Núcleo de Informações Estratégicas e outros três auditores em atividades de análise de licitações com experiência em casos de corrupção.

A condução dos trabalhos foi essencialmente com entrevistas informais com periodicidade a depender da necessidade do caso, incluiu diversas tarefas como construção de atributos, integração de bases, tratamento de dados, interpretação de resultados entre outras

atividades. As sugestões desse corpo técnico foram aplicadas nas quatro dimensões definidas e na escolha da técnica de aprendizado de máquina.

4.1. Entendimento do Negócio

Para o entendimento do negócio foram consultados os especialistas em diversas áreas no Tribunal de Contas de forma a permitir a definição dos atributos para formação do conjunto de dados, bem como da literatura pesquisada no capítulo anterior.

4.1.1. Combate a corrupção

Entre o rol de competências do Tribunal de Contas do Distrito Federal está a fiscalização de pessoal e combate aos desvios de recursos públicos, conforme se depreende da Lei Orgânica do Distrito Federal transcrita em parte a seguir:

> Art. 78. O controle externo, a cargo da Câmara Legislativa, será exercido com auxílio do Tribunal de Contas do Distrito Federal, ao qual compete:
>
> (...) II – julgar as contas:
>
> a) dos administradores e demais responsáveis por dinheiros, bens e valores da administração direta e indireta ou que estejam sob sua responsabilidade, incluídos os das fundações e sociedades instituídas ou mantidas pelo Poder Público do Distrito Federal, bem como daqueles que derem causa a perda, extravio ou outra irregularidade de que resulte prejuízo ao erário;
>
> (...) V – realizar, por iniciativa própria, da Câmara Legislativa ou de alguma de suas comissões técnicas ou de inquérito, inspeções e auditorias de natureza contábil, financeira, orçamentária, operacional e patrimonial, nas unidades administrativas dos Poderes Executivo e Legislativo do Distrito Federal:
>
> (...) IX – aplicar aos responsáveis, em caso de ilegalidade de despesa ou irregularidade de contas, as sanções previstas em lei, a qual estabelecerá, entre outras cominações, multa proporcional ao dano causado ao erário;
>
> (...) XIV – apreciar e apurar denúncias sobre irregularidades e ilegalidades dos atos sujeitos a seu controle.

A Lei Complementar Distrital n.º 01/1994 dispõe sobre a Lei Orgânica do Tribunal de Contas do Distrito Federal [16] e estabelece de forma transparente a competência do Tribunal na fiscalização de pessoal quanto aos atos e desvios de recursos, conforme se depreende dos artigos a seguir:

> Art. 9º Diante da omissão no dever de prestar contas, da não comprovação da aplicação dos recursos repassados pelo Distrito Federal, na forma prevista no inciso VI do art. 6º desta Lei Complementar, da ocorrência de desfalque ou desvio de dinheiros, bens ou valores públicos, ou, ainda, da prática de qualquer ato ilegal, ilegítimo ou antieconômico de que resulte dano ao Erário, a autoridade administrativa competente, sob pena de responsabilidade solidária, deverá imediatamente adotar providências, com vista à instauração de tomada de contas especial, para apuração dos fatos, identificação dos responsáveis e quantificação do dano.
>
> Art. 17. As contas serão julgadas:
>
> (...) d) desfalque ou desvio de dinheiros, bens ou valores públicos.
>
> (...) Art. 46. Ao exercer a fiscalização, se configurada a ocorrência de desfalque, desvio de bens ou outra irregularidade de que resulte dano ao Erário, o Tribunal ordenará, desde logo, a conversão do processo em tomada de contas especial, salvo a hipótese prevista no art. 84 desta Lei Complementar.

Como faz parte das atribuições do TCDF os especialistas são servidores do corpo técnico que auxiliaram para formação de conhecimento e integração das bases de dados.

4.2. Entendimento dos Dados

Para o processo de mineração de dados, buscou-se entre as bases de dados disponíveis as que poderiam apresentar os atributos identificados na literatura.

Um trabalho de dissertação apresentada neste PPCA de título “Modelos Preditivos para Avaliação de Risco de Corrupção de Servidores Públicos Federais [48]”, referenciado no item 2.2, apresentou considerável contribuição na definição de atributos e foi relacionado em todas as quatro as dimensões de atributos.

Alguns dos atributos que representariam o patrimônio e consumo dos servidores não foram possíveis de incluir para a presente investigação. Esses atributos poderiam inclusive permitir identificar variações de patrimônio ou até consumo incompatível com a renda dos servidores, estas bases seriam as que tratam de dados de IPTU, IPVA, consumo de água e de energia elétrica.

No entanto, foi possível obter a maior parte dos atributos identificados na literatura a partir de oito bases de dados descritas a seguir:

- ✓ Corregedoria Geral do Distrito Federal (Portal da Transparência DF);
- ✓ SIGRH - Sistema Integrado de Gestão de Recursos;

- ✓ SIAPE - Sistema Integrado de Administração de Recursos Humanos;
- ✓ TCDF com informações de inabilitados para o exercício de cargo em comissão ou função de confiança no âmbito da Administração Pública do Distrito Federal;
- ✓ Entidades Privadas sem Fins Lucrativos Impedidas de contratar com a Administração Pública (CEPIM) Controladoria Geral da União;
- ✓ Cadastro de Empresas Inidôneas e Suspensas (CEIS) Controladoria Geral da União;
- ✓ Tribunal Superior Eleitoral (Dados eleitorais);
- ✓ Secretaria da Receita Federal (dados cadastrais de pessoa física e jurídica)

Os dados obtidos dessas bases de dados estão delineados por seus atributos nos tópicos adiante que tratam o tema por dimensões temáticas (Corrupção, Funcional, Política e Vínculos Societários).

4.2.1. Dimensão de Corrupção

A dimensão corrupção trata de atributos pertinentes a punibilidades aplicadas a servidores ou as empresas que possam ter servidores públicos como sócios. A seguir apresentam-se os atributos selecionados nas bases disponíveis.

TARGET_CORRUPCAO é o atributo que indica a expulsão de servidor do GDF fundamentada na Lei 8.429/92 [4] por 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. Os dados estão disponíveis publicamente, ou seja, é dado aberto que pode ser obtido em <www.transparência.df.gov.br>.

CEIS é o atributo obtido do Cadastro de Empresas Inidôneas e Suspensas e apresenta o detalhamento das sanções vigentes para pessoa física ou jurídica com restrição ao direito de participar em licitações ou de celebrar contratos com a Administração Pública (Federal, estadual, distrital ou municipal), está disponível em http://www.portaltransparencia.gov.br, é mantida pela Controladoria Geral da União – CGU e busca cumprir determinações da Lei nº 12.846/2013 (Lei Anticorrupção).

O atributo CEPIM foi obtido do Cadastro de Entidades Privadas Sem Fins Lucrativos Impedidas (CEPIM) que apresenta a relação de entidades privadas sem fins lucrativos que estão impedidas de celebrar novos convênios, contratos de repasse ou termos de parceria com a

Administração Pública Federal, em função de irregularidades não resolvidas em convênios, contratos de repasse ou termos de parceria firmados anteriormente.

O TCDF mantém rol de pessoas inabilitadas para o exercício de cargo em comissão ou função de confiança no âmbito da Administração Pública do Distrito Federal pelo período de até oito anos, em decorrência de irregularidades graves constatadas pelo TCDF, nos termos do art. 60 da Lei Complementar n.º 01/1990. O atributo que apresenta essa característica foi denominado InabilitadosTCDF.

4.2.2. Dimensão Funcional

As bases de dados utilizadas nessa dimensão foram o Sistema Integrado de Gestão de Recursos (SIGRH) e o Sistema Integrado de Administração de Recursos Humanos (SIAPE). São duas bases diferentes para pagamento dos servidores do Governo do Distrito Federal que separam os servidores da Segurança Pública (SIAPE) dos outros servidores (Educação, saúde e demais áreas).

Os atributos criados nessa dimensão foram:

SalBrutoF – salário bruto do servidor que incluiu o salário recebido por qualquer uma das bases (SIGRH e SIAPE) ou a soma dos salários para o caso de servidores que acumulam cargos públicos conforme permissivo previsto na Constituição Federal.

SaliqF – salário líquido do servidor com diversos descontos e obtido de forma análoga ao Salário Bruto (bases SIGRH e SIAPE).

Os próximos atributos foram criados a partir de variáveis categóricas que expressavam o cargo ou função do servidor durante a vida funcional, a partir dessa informação elaborou-se um atributo de contagem para expressar quantos cargos ou funções o servidor ocupou na administração pública do DF até a data da apuração da pesquisa (janeiro/2020).

QTCG_SIGRH (numérica) – quantidade de cargos que o servidor ocupou até a data da apuração da investigação no SIGRH (Servidores, exceto da Segurança Pública) determinado somente com a base SIGRH.

QTCG_SIAPE (numérica)– quantidade de cargos que o servidor ocupou na Segurança Pública até a data da apuração da investigação no SIAPE (Segurança Pública, SIAPE)

QTCG_SIGRHSIAPE (numérica)– quantidade de cargos que o servidor ocupou até a data da apuração da investigação nas duas bases (SIGRH e SIAPE).

QTFCSIGRH (numérica) – quantidade de funções que o servidor ocupou até a data da apuração da investigação no SIGRH (Servidores, exceto da Segurança Pública, SIGRH)

QTFCSIAPE (numérica) – quantidade de funções que o servidor ocupou até a data da apuração da investigação no SIAPE (Segurança Pública SIAPE)

QTFCSIGRHSIAPE (numérica) - quantidade de funções que o servidor ocupou até a data da apuração da investigação nas duas bases (SIGRH e SIAPE).

4.2.3. Dimensão Política

Atributos dessa dimensão foram obtidos exclusivamente das bases disponíveis no Tribunal Superior Eleitoral, uma base de dados aberta que apresenta diversas características dos candidatos a cargo eletivo. Dessa base foram identificados os servidores públicos do DF candidatos a cargo eletivo nas eleições de 2014 e obtidos os atributos de informações eleitorais, conforme descrito a seguir:

CODIGO_CARGO – Tipo de cargo eleitoral que o servidor disputou (presidente ou vice, governador ou vice, prefeito, senador, vereador, deputado federal, estadual ou distrital);

NUMERO_PARTIDO – o número do partido em que o servidor estava inscrito para eleição;

COD_SITUACAO_CANDIDATURA – é a descrição da situação do registro de candidatura do candidato que pode assumir os valores 'Apto' (candidato apto para ir para urna); 'Inapto' (candidato inapto para ir para urna); 'Cadastrado' (registro de candidatura realizado, mas ainda não julgado pela instância eleitoral);

COD_GRAU_INSTRUCAO – O grau de instrução do candidato pode ser definido como: não divulgável, lê e escreve, ensino fundamental incompleto ou completo, ensino médio incompleto ou completo e ensino superior incompleto ou completo.

CODIGO_ESTADO_CIVIL – É a situação de Estado civil do candidato servidor público: solteiro(a), casado(a), não divulgável, viúvo(a), separado(a) judicialmente ou divorciado(a).

COD_SIT_TOT_TURNO – Este atributo identifica a situação de totalização do candidato no turno que pode ser: eleito, eleito por média, eleito por quociente eleitoral, não eleito, suplente ou nulo.

Todas esses atributos são categóricos, ou seja, representam categorias que podem representar gradação ou apenas classificação.

4.2.4. Dimensão de Vínculos Societários

Nessa dimensão descrevem-se os atributos de vínculos societários que podem compor como característica de servidores públicos do Distrito Federal (servidores/aposentados). A base de dados de vínculos societários é a base da Receita Federal do Brasil, que registra a maior parte dos atributos como dados abertos e disponíveis em <www.receita.economia.gov.br>.

A seguir, lista-se os atributos utilizados na pesquisa:

TIPO_SOCIO – esse atributo descreve qual o tipo de sócio o servidor está cadastrado na empresa em que faz parte;

CNAE_FISCAL – o código da atividade principal da empresa em que o servidor é sócio;

QTD_CNAES – quantidade de atividades secundárias cadastradas da empresa em que o servidor é sócio;

COD_NATUREZA_JURIDICA – natureza jurídica da empresa em que o servidor é sócio que pode ser em diversas denominações como: Sociedade de Economia Mista, Sociedade Anônima Aberta ou fechada, Sociedade Empresária Limitada, comandita, ou por ações, além de outras;

COD_SITUACAO_CADASTRAL – situação cadastral da empresa em que o servidor é sócio, entre as alternativas possíveis têm-se ativa, nula, suspensa, inapta ou baixada;

COD_PORTE_EMPRESA – Este atributo descreve o porte da empresa que pode ser Microempreendedor Individual (MEI), Microempresa (ME), Empresa de Pequeno Porte (EPP), de médio porte ou grande porte a depender do faturamento anual bruto da matriz e suas filiais, ou seja, o faturamento bruto global definido na legislação tributária;

COD_OPCAO_SIMPLES – este atributo informa se a empresa optou pelo sistema de tributação simplificado – Simples Nacional – que objetiva auxiliar as empresas de micro e pequeno porte em relação ao pagamento de tributos;

COD_QUALIFICACAO_SOCIO – trata do tipo de qualificação do sócio que pode ser diretor, presidente, administrador, conselheiro de administração, sócio, sócio menor, incapaz ou relativamente incapaz, e outras denominações;

DiasNaSociedade – este atributo informa a quantidade de dias em que o servidor é sócio na empresa até a data da apuração dessa pesquisa;

PERCENTUAL_CAPITAL_SOCIAL – percentual do capital social que o servidor sócio apresenta na data de apuração dessa pesquisa.

Apesar dos atributos CEIS e CEPIM apresentarem relação com vínculos empregatícios foram tratados na dimensão corrupção devido à característica de punibilidade.

4.3. Preparação ou pré-processamento de Dados

A preparação dos dados é a etapa em que os dados devem ser processados e preparados de forma que possam evidenciar o entendimento do negócio. A integração de diversas fontes de dados pode ser um desafio porque, em geral, os dados são oriundos de fontes de sistemas transacionais, ou medições, ou também de situações reais e o conjunto de dados obtidos deve convergir para o entendimento do negócio.

Para isso, algumas subetapas são realizadas de forma a gerar dados tratados e adequados ao processo de modelagem. Estas subetapas são: limpeza de dados, construção de atributos, análise de variância, correlação e separação de dados.

4.3.1. Limpeza de dados

Essa subetapa pretende amenizar dois problemas importantes de processos de aquisição de dados: existência de valores ausentes (*missing values*) e existência de valores ruidosos (*noise values*).

A ausência de valores ocorre quando para os atributos de um conjunto de dados não se apresenta valor determinado para alguns exemplares ou quando um conjunto de dados não possui valores para um atributo de interesse ou ainda apresenta valores agregados em relação a esse atributo.

Como solução para ausência de valores é possível a remoção dos exemplares com essa característica, preenchimento manual de valores ou preenchimento automático.

Os valores ruidosos referem-se às modificações dos valores originais e que, portanto, consistem em erros de medidas ou em valores consideravelmente diferentes da maioria dos outros valores do conjunto de dados, conhecidos como *outliers*. Como exemplo pode-se citar casos que deveriam ser positivos e ocorrem valores negativos ou uma mudança de comportamento dos valores de um atributo sem explicação.

Para solução de valores ruidosos, tem-se a inspeção com correção manual ou identificação e limpeza automática implementada por algoritmos que suavizam ou anulam ruídos.

4.3.2. Construção de Atributos

A construção de atributos permite elaborar atributos que possam gerar informações relevantes segundo o entendimento do negócio a partir dos dados originais. Para o processo de mineração alguns atributos devem ser modificados de forma a permitir o uso de algoritmos de aprendizagem de máquina.

Na presente pesquisa foi empregado o uso de transformação de variáveis categóricas, ou seja, variáveis que descrevem categorias ou classificações, para variáveis binárias, variáveis com valor 0 ou 1 que expressam existência ou ausência do atributo binário. Esse procedimento também é conhecido como aplicação para *dummy variables*.

Outra perspectiva de construção de atributos utilizada foi a transformação de atributos categóricos em atributo de contagem. Esse procedimento foi realizado porque o atributo ao expressar quantidade apresenta significado no contexto de entendimento de negócio, enquanto o valor categórico não expressa benefício no contexto da investigação.

4.3.3. Análise de Variância e Correlação

A análise de variância e de correlação de dados destina-se a selecionar atributos que não geram informação quando aplicado aos algoritmos de aprendizagem de máquina, prejudicam ou não contribuem para geração de informação.

Os atributos com variância zero, ou seja, atributo que apresenta frequência zero para uma mesma classe devem ser excluídos pois não representam um atributo que contribua para o modelo, como resultado, o algoritmo de regressão logística irá retornar coeficiente nulo ou infinito prejudicando a possibilidade de convergência do algoritmo. De forma análoga, ocorre para os atributos com variância quase zero.

A análise de correlação evidencia quais atributos apresentam correlação estatística entre si. Essa avaliação permite selecionar atributos de alta correlação, ou seja, atributos que tenham alta dependência ou associação indicando mesma informação. O efeito dessa característica é a presença de atributos redundantes capazes de induzir ao erro o modelo preditivo.

A exclusão dos atributos com variância zero ou quase zero e dos atributos com alta correlação auxilia na redução de dimensionalidade, como efeito dessa redução temos a redução do custo de processamento e possível incremento na precisão do classificador.

4.3.4. Separação de dados

Após a redução da dimensionalidade com exclusão de atributos que não contribuem para o modelo preditivo inicia-se a separação de dados em teste e treino que foi realizada na proporção de 10% dos dados para teste e 90% para treino com rotação desse grupo de dados até completar um ciclo completo, um procedimento conhecido como *Cross-validation*.

Dada a complexidade do desbalanceamento dos dados em situação extrema presente nessa investigação, aplicou-se uma funcionalidade presente na programação Python para forçar a distribuição aleatória com mesma proporção de casos das classes da variável dependente em treino e teste, ou seja, manter a proporção de desbalanceamento do conjunto de dados no conjunto de teste e de treino.

4.4. Modelagem

A fase de modelagem ocorre com a criação de modelo com uso dos algoritmos aplicados nos dados. Para a presente pesquisa o desbalanceamento representa situação em que as técnicas empregadas para mineração de dados devem ser aplicadas de forma que o algoritmo de aprendizado não apresente viés da classe majoritária, que não é a classe de interesse. Para isso, as opções descritas na literatura e presentes no tópico 2.4 são [70]:

- ✓ solução com tratamento de dados, balanceamento dos dados com técnicas de reamostragem para adequado treinamento dos algoritmos com uso de SMOTE ou;
- ✓ solução com algoritmo otimizado com características específicas para cenários de desbalanceamento de classes.

Essas abordagens serão tratadas no próximo capítulo.

Após discussão com especialistas do TCDF com os resultados das técnicas de mineração, foi definido a aplicação da regressão logística nessa pesquisa, pois a variável de

interesse é dicotômica (corrupção ou não corrupção); o algoritmo apresenta fácil entendimento de uso e interpretação de resultado em relação a outras técnicas como aprendizado aprofundo (*deep learning*) e, ainda, pela falta de maturidade da instituição no uso de inteligência artificial, como outros órgãos da Administração Pública, que poderia dificultar a implementação e uso das técnicas com menor transparência.

5. Resultados

Os resultados alcançados serão apresentados em relação aos objetivos específicos definidos na seção 1.4.2.

5.1. Identificar os fatores de riscos relativos à corrupção de servidores públicos

A pesquisa bibliográfica obteve alguns fatores de risco que foram utilizados, com a opinião de especialistas, para formação dos atributos para composição do conjunto de dados para avaliação preditiva.

A pesquisa apresentou os fatores divididos em quatro dimensões: Corrupção, Funcional, Política e de Vínculos Societários. Os principais trabalhos são delineados a seguir.

Dimensão corrupção

No estudo de Hanna [33], dois aspectos importantes foram detectados como predisposição a corrupção, servidores que burlaram provas e servidores que apresentam propensão a desonestidade. A primeira característica não nos permitiu elaborar consultas em bancos de dados por não existirem informações desse gênero no Brasil, no entanto, a segunda característica foi utilizada com obtenção de antecedentes de punibilidade registrados de servidores.

O estudo de Lassou [34] informa dos problemas de transparência, de responsabilização e de contabilidade governamental como fatores de corrupção, nesse aspecto, a possibilidade de avaliação de risco de corruptibilidade de servidores públicos pode ser considerada relevante para órgãos considerados com baixa transparência, permitindo a ação como fator ambiental facilitador.

Para Padula & Albuquerque [7] os funcionários públicos apresentam a tendência de manter e preservar as dificuldades das atividades realizadas para garantir o recebimento ilícito. Os prazos longos nos processos licitatórios e renovações emergenciais podem ser caracterizados nesse contexto.

Dimensão Funcional

O estudo de Gans-Morse [10] aponta que a corrupção emerge de situações em que servidores públicos apresentam salários abaixo do mínimo básico ou em casos de cortes

salariais em consequência de medidas austeras, ou crises do Estado. Também acrescenta que os subornos são considerados moralmente aceitáveis quando os salários dos servidores estão abaixo do limiar de pobreza. Outro aspecto relevante nesse cenário é que o aumento do salário não necessariamente reduz a corrupção presente. Desse estudo pode-se estabelecer relação do tema com a remuneração e variações.

Outro aspecto relevante no processo de seleção de servidores públicos na Ásia foi apontado por Poocharoen [37], que informa da dificuldade de tornar o sistema justo e livre de corrupção pela escolha para nomeação de servidores públicos com a melhor qualificação sem a existência de clientelismo. Essa relação pode ser identificada pela característica de um servidor ser ou não efetivo, ou até nomeado a cargo público.

Kuotsai [40] observou um papel limitado da alta remuneração no controle da corrupção no serviço público da China e sustenta que a prevenção da corrupção requer esforços conjuntos em várias áreas da gestão de recursos humanos.

Dimensão Política

Bersch & Matthew [46] apresentam como conclusão a relação da dominação partidária de nomeações políticas que reduz a capacidade de prestação de serviço adequado e relaciona com a capacidade de combater efetivamente a corrupção.

A pesquisa de Meyer-Sahling & Mikkelsen [47] evidencia que quanto maior a politização maior é o nível de corrupção, enquanto maior o mérito do recrutamento menor é a corrupção associada.

Dessas conclusões, pode-se estabelecer relação de diferenciação no risco de corrupção para servidores concursados (efetivos) e servidores que não apresentam vínculo com a administração pública. Quanto a politização, cabe agregar os conhecimentos de outros estudos consultados para estabelecer a forma de identificação de risco de corruptibilidade com os dados presentes em banco de dados.

Como forma de combate a corrupção, Moro [6] entende que deveria haver uma regulação legal estrita na contribuição das empresas para as campanhas eleitorais e cita como exemplo o caso de empresa com contrato com o governo com doações a campanha eleitoral.

Dimensão de vínculos societários

Em Carvalho [48] os autores identificaram como fator de risco os sócios das empresas que são servidores públicos, além deste, outros atributos foram eleitos por especialistas para avaliação na elaboração do modelo.

5.2. Identificar as técnicas de mineração de dados para o contexto de corrupção e fraude

Para este objetivo específico utilizou-se da pesquisa bibliográfica com o termo "*data mining corruption*" que retornou 78 artigos da *Web of Science* e 130 artigos da *Scopus* restrito aos últimos dez anos. Muitos artigos foram descartados por tratarem de outras áreas como corrupção de imagens, de memória ou de dados e ainda, de empresas de mineração de pedras.

Identificaram-se diversas técnicas de aprendizagem de máquina para a área de pesquisa, entre elas, cabe citar: análise de desvios ou anomalias [62], regras de associação [58][61], *Naive Bayes*[49][52][59][64], *Clustering* [55][58][61][62], *k-means* [55][65], *Support Vetor Machine* [52][57], *Random Forest* [52], algoritmos de *deep learning* como redes neurais [56] [57] [60] [62] e mineração na web [66] [67]. Estes algoritmos e as referências bibliográficas estão sintetizados na figura 3.5 da seção 2.2.

A leitura dos artigos obtidos nessa pesquisa de técnicas indicou como relevante a investigação do tema de desbalanceamento de dados. Isso ocorre porque a característica de casos de fraude e corrupção no dia a dia das transações reais é apresentar pouca expressividade de casos, algo inferior ou próximo a 1% das transações.

Desse modo, foi pertinente a pesquisa de como lidar nesse contexto de desbalanceamento de dados, pois os algoritmos de aprendizado de máquina tendem a classificar para as classes de dados majoritárias e o caso de interesse nessa investigação é a classe de dados minoritária.

A técnica *synthetic minority oversampling technique (SMOTE)* [23] foi elaborada para lidar com o desbalanceamento de classes para casos de fraudes, publicada em 2002 e com crescente ampliação de estudos ao longo dos anos, conforme se denota no quantitativo de publicações disponível na *Web of Science*, figura 3.6 da seção 2.4.

5.3. Elaboração do modelo preditivo e interpretação

Para elaboração do modelo foram utilizadas oito bases de dados para compor o conjunto de dados para mineração, conforme quadro a seguir por dimensão.

DIMENSÃO CORRUPÇÃO
Corregedoria Geral do Distrito Federal
Entidades Privadas sem Fins Lucrativos Impedidas de contratar com a Administração Pública (CEPIM)
Cadastro de Empresas Inidôneas e Suspensas (CEIS)
TCDF inabilitados para o exercício de cargo em comissão ou função de confiança no âmbito da Administração Pública do DF
DIMENSÃO FUNCIONAL
SIGRH - Sistema Integrado de Gestão de Recursos
SIAPE - Sistema Integrado de Administração de Recursos Humanos
DIMENSÃO POLÍTICA
Tribunal Superior Eleitoral (Dados eleitorais)
DIMENSÃO SOCIETÁRIA
Secretaria da Receita Federal (dados cadastrais de pessoa física e jurídica)

Figura 5.1 – Fonte de dados por Dimensão de pesquisa.

As bases de dados são oriundas do Governo Federal e Distrital, foram importadas para uso no ambiente SAS e todo o tratamento de limpeza de dados (3.3.1) foi realizado pelo software *SAS Enterprise Guide* enquanto as demais etapas na IDE Anaconda com uso das linguagens Python e R.

No item 3.2 – Entendimento dos dados – listou-se todos os atributos obtidos dessas bases de dados. Os atributos da dimensão Corrupção são binários, portanto, não houve nenhuma transformação para aplicação aos algoritmos.

Cabe informar que a variável dependente TARGET_CORRUPCAO corresponde ao atributo que informa o servidor que foi expulso por corrupção, ato de improbidade administrativa nos termos da Lei º 8429/92 [4], esse atributo assume valor 0 (“não corrupto”) ou 1 (“corrupto”).

O conjunto de dados corresponde a população de servidores do Governo do Distrito Federal, incluindo os aposentados e considerando que um servidor pode ser sócio de mais de uma empresa totalizando 303.036 registros na data de apuração da pesquisa (janeiro de 2020).

Reiterando a informação do tópico 2.4, o aprendizado de máquina aplicado em conjunto de dados com desbalanceamento extremo requer tratamento específico para que o algoritmo possa obter os padrões desejados da classe minoritária sem gerar viés da classe dominante.

O atributo que é variável independente dessa investigação (TARGET_CORRUPCAO) apresenta na classe de interesse 428 registros e na classe dominante 302.608 registros, uma relação que guarda a proporção de 1:707, em termos percentuais 0,14% da classe de interesse em relação à população. As possíveis formas de tratar este cenário serão abordadas nesse capítulo.

As bases da dimensão Funcional (SIGRH e SIAPE) foram agregadas, pois um mesmo servidor do DF pode constar nas duas bases devido a possibilidade prevista na Constituição Federal de acumulação de cargos públicos.

Para essas bases, alguns atributos categóricos foram transformados em numéricos de contagem para refletir o entendimento de negócio. Isso ocorreu para cargos e funções das duas bases em que a informação original (categórica) não apresenta significado no contexto de corrupção, mas a quantidade de cargos ou funções que o servidor assumiu durante a vida funcional até a data da apuração da pesquisa reflete significado.

Os atributos da dimensão política são originalmente categóricos e para aplicação aos algoritmos foram transformados em atributos binários, conforme explicitado no item 3.3.2 – Construção de Atributos. Do mesmo modo procedeu-se com os atributos categóricos da dimensão Societária.

Inicialmente o conjunto de dados compreendia 28 atributos (numéricos e categóricos) que após transformações necessárias das categóricas resultaram em 11 numéricas e 1.116 binárias.

E como última transformação, de forma a evitar viés dos atributos numéricos no algoritmo, esses atributos numéricos foram normalizados, ou seja, cada valor foi subtraído do menor valor do atributo e dividido pela amplitude (maior valor subtraído do menor valor do atributo) resultando em valores entre zero e um.

Após as subetapas de limpeza de dados (3.3.1) e construção de atributos (3.3.3), procedeu-se a análise de variância e correlação (3.3.3).

Quanto a avaliação de correlação entre variáveis, identificou-se e excluiu-se quatro atributos com correlação de Pearson acima de 0,9 e sete atributos com correlação entre 0,8 e 0,9. Com relação a variância dos atributos, CEPIM apresentou valor nulo, 294 atributos apresentaram variância menor que 0,00001 e 421 atributos registraram variância entre 0,00001 e 0,0001. Esses atributos foram excluídos reduzindo a dimensionalidade de 1.127 atributos para 397 atributos.

A literatura indica como parâmetro adequado de avaliação de modelagem em cenário de classes extremamente desbalanceadas a área sobre a curva ROC (*Receiver operating characteristic*) [71]. A curva ROC é um gráfico que ilustra a capacidade de diagnóstico de um sistema classificador binário conforme seu limite de discriminação é variado.

Diversas experimentações de modelagem com regressão logística foram realizadas de forma a obter maximização da área sobre a curva ROC, transformações diferenciadas em diversos atributos, exclusões massivas de atributos e os resultados mostraram-se de baixa qualidade de modelagem, valores de área da curva ROC inferiores a 0,6.

Segundo Brownlee [71], o desbalanceamento extremo é um desafio para modelagem que requer técnicas especializadas. Nessa investigação abordou-se as duas possibilidades de tratamento de desbalanceamento de dados.

A primeira linha de ação foi o tratamento de dados de forma tornar as classes balanceadas. Na literatura a técnica com maior número de citações foi SMOTE e suas variações, conforme explicitado no item 2.4.

Os valores obtidos como aplicação da técnica SMOTE e variações estão resumidos na tabela a seguir.

Tabela 5.1 – Resultados de área sobre a curva ROC com variações de SMOTE

TÉCNICAS	ÁREA ROC	PACOTES
SMOTE e random undersampling	0,658	from imblearn.over_sampling & from imblearn.under_sampling (Python)
SMOTE and Tomek Links sampling	0,534	imblearn.combine & imblearn.under_sampling (Python)
SMOTEENN	0,601	imblearn.combine (Python)
SMOTE	0,422	SmoteFamily (R)
DBSMOTE	0,534	SmoteFamily (R)
ADAS	0,548	SmoteFamily (R)
ANS	0,534	SmoteFamily (R)
SLS	0,491	SmoteFamily (R)

Fonte: SmoteFamily disponível https://cran.r-project.org/web/packages/smotefamily/smotefamily.pdf

Os resultados obtidos com aplicação de tratamento nos dados com reamostragem e a técnica SMOTE e variações geraram resultados da área de curva ROC de baixa qualidade, e que se mostram inadequados para criação de modelo para os dados da presente investigação.

Desse modo, optou-se pela experimentação de exclusões de atributos de forma a obter a maximização da área sobre a curva ROC.

Na segunda linha de ação, utilização de algoritmos de regressão com características específicas para lidar com o desbalanceamento obteve-se a apuração da performance com relação a área da curva ROC com os procedimentos de normalização das variáveis numéricas, exclusões de atributos com correlação superior a 0,8 e variância inferior a 0,0001 que apresentou o melhor resultado de modelagem segundo as melhores práticas de parametrizações de algoritmos com aplicação de regressão logística.

A seguir apresenta-se quadro com resultados da área sobre a curva ROC com os passos de exclusões de atributos.

Tabela 5.2 – Resultados de área sobre a curva ROC com exclusão de atributos

	TIPOS DE ATRIBUTOS				**MODELOS**	
	numéricas	**categóricas**	**binárias**	**Total**	**I**	**II**
Atributos originais	11	13	4	28	-	-
1-Transformação dos atributos categóricos em binários	11	0	1116	1127	0.692	0.642
2-Exclusões de atributos com:						
variância zero e correlação > 0.9	9	0	1111	1120	0.692	0.644
3-Correlação entre 0,8 e 0,9	8	0	1104	1112	0.692	0.643
4-Variância <0.00001	8		810	818	0.692	0.643
5-Variância entre 0.00001 e 0.0001	8	0	389	397	0.692	0.647

Obs: Os atributos excluídos estão disponíveis em https://github.com/marcelovasc/MRC com o código em Python além de todos os procedimentos realizados.

Os gráficos a seguir representam a área sobre a curva ROC determinada com os passos de exclusões de atributos descritos na tabela anterior.

O acompanhamento de resultados da área sobre a curva ROC com exclusão de variáveis permitiu identificar a melhor possibilidade de exclusão de atributos para redução de dimensionalidade sem afetar de forma significativa a performance do modelo a ser gerado.

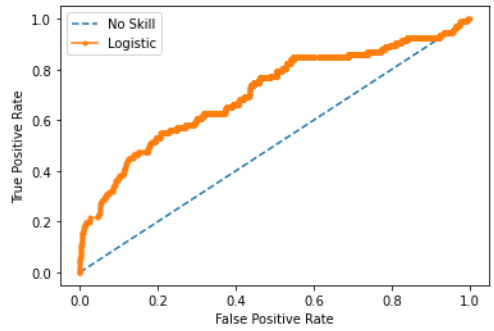
Exclusões de atributos com correlação menores que 0,8 e variâncias maiores que 0,0001 apresentaram redução de área e, portanto, foram alternativas desconsideradas e optou-se por manter os patamares alcançados.

Os dois modelos foram elaborados com regressão logística ponderada (LOGIT) por meio da biblioteca *Scikit-Learn* implementados em *Python*.

Os valores apresentados de cada etapa representam médias da área sobre a curva ROC calculada a partir de estratificação de validação cruzada (*cross validation*) com 10 pastas ou partições e refeitas por três vezes de forma a representar valor mais adequado para medição.

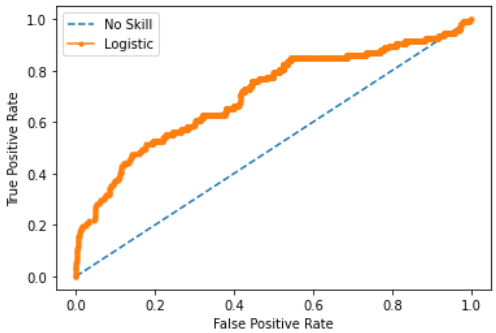
Cabe ressaltar que também foi utilizado recurso da biblioteca para garantir que nas partições da validação cruzada contenham amostras proporcionais da classe minoritária que é a classe de interesse da pesquisa, aspecto relevante no trato de desbalanceamento extremo [71].

1-Dados transformados



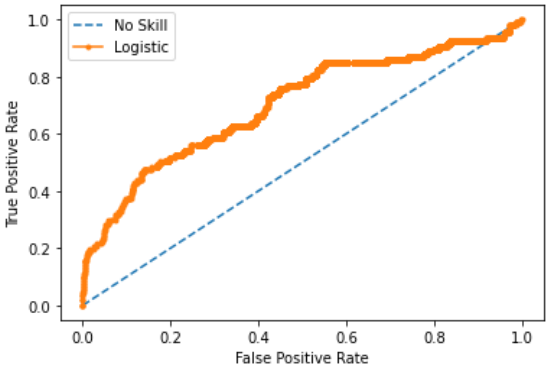
Área sobre ROC = 0.7069536554740817

2-variância zero e correlação > 0.9



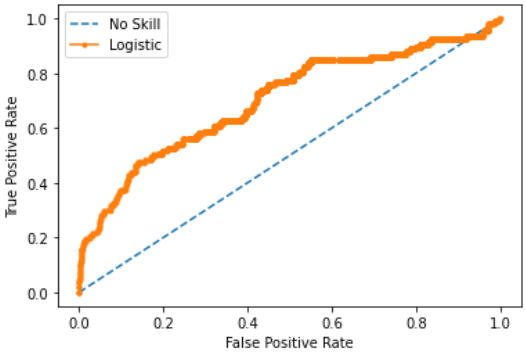
Área sobre ROC = 0.7065263422133121

3-Correlação acima de 0,8



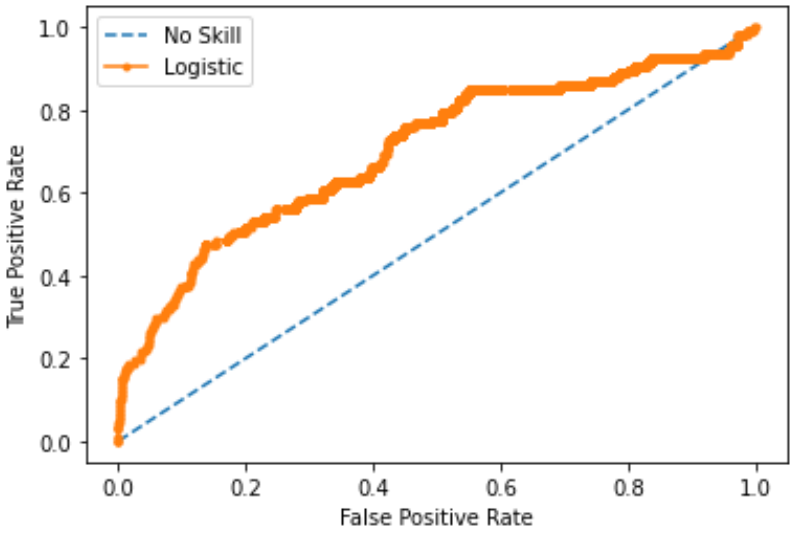
Área sobre ROC = 0.7037229250908366

4-Variância <0.00001



Área sobre ROC = 0.7036973530049795

5-Variância entre 0.00001 e 0.0001



Área sobre ROC = 0.7034893172920174

Figura 5.2 – Curva ROC com exclusão de atributos.

No modelo I foi definido como hiperparâmetro a relação de peso 0.0014, que é a proporção de casos da classe minoritária em relação a majoritária.

No modelo II foi utilizada implementação heurística de melhores práticas na ponderação de classes disponível na biblioteca e que determina o peso atribuído como a divisão

da quantidade de populacional pelo produto do número de classes pela quantidade populacional da classe majoritária.

Entre os dois modelos a atribuição fixada de peso 0.0014 apresentou melhor resultado. A partir desse resultado de modelagem (I), estabeleceram-se os coeficientes dos atributos na regressão logística.

No contexto da teoria de gerenciamento de risco, esses coeficientes da regressão logística representam o fator de risco de cada atributo em relação à possibilidade de corrupção, pois a representatividade numérica desse coeficiente está relacionada a variável resposta ou atributo alvo da investigação.

Outra medida para redução de dimensionalidade foi a exclusão de variáveis que possam ser considerados combinações lineares entre si, para isso verificou-se no modelo de regressão logística os preditores com *p-value* superior a 0,05. Excluindo-se esses atributos o modelo gerado restou com sessenta e oito atributos.

A lista de todos os coeficientes obtidos na regressão logística para os atributos está disponível no anexo I desta dissertação.

A seguir apresenta-se os principais valores de coeficientes da regressão logística em ordem decrescente e o exponencial neperiano do coeficiente que é conhecido como *odds ratio*.

Tabela 4.3 – Coeficientes dos atributos da regressão logística

N	ATRIBUTO	Coeficiente (x)	e(x)	N	ATRIBUTO	Coeficiente (x)	e(x)
1	QTFCSIGRHe	1.6068	4.9870	35	CNAE8219999	0.3188	1.3755
2	CEIS	1.1246	3.0791	36	CNAE4761002	0.3138	1.3686
3	CNAE8531700	0.8275	2.2876	37	OPCAO_SIMPLES0	0.2995	1.3491
4	CNAE8640202	0.7255	2.0658	38	CNAE9493600	0.2939	1.3417
5	QTCG_SIAPEe	0.7099	2.0338	39	CNAE4930203	0.2933	1.3408
6	CNAE6911701	0.6735	1.9610	40	CNAE8299799	0.2881	1.3339
7	CNAE4754701	0.6575	1.9299	41	CNAE8011101	0.2625	1.3001
8	PERCENTUAL_CAPITAL_SOCIAL	0.6256	1.8693	42	CNAE4756300	0.2598	1.2967
9	CNAE4789001	0.6154	1.8503	43	QUALIFSOCIO22	0.2429	1.2749
10	OPCAO_SIMPLES6	0.5941	1.8115	44	CNAE6424703	0.2421	1.2739
11	CNAE7490199	0.5658	1.7609	45	NATURJURID2305	0.2156	1.2405
12	CNAE5811500	0.5618	1.7539	46	CargoEleitoral8	0.2011	1.2227
13	CNAE6920601	0.5398	1.7156	47	CNAE7490101	0.1925	1.2123
14	EstadoCivil1	0.5336	1.7051	48	CNAE4520005	0.1908	1.2103
15	GrauInst6	0.5316	1.7017	49	CNAE4530701	0.1773	1.1940
16	QUALIFSOCIO18	0.4945	1.6397	50	PORTEEMPRESA05	0.1637	1.1778
17	CNAE4530703	0.4824	1.6199	51	QUALIFSOCIO65	0.0948	1.0994
18	CNAE5611202	0.4755	1.6087	52	TPSOCIO2	0.0121	1.0121
19	QUALIFSOCIO49	0.4664	1.5943	53	SITUACADASTRAL08	(0.0182)	0.9819

20	CNAE8593700	0.4534	1.5737	54	CNAE4693100	(0.0228)	0.9774
21	CNAE6822600	0.4521	1.5716	55	CNAE4789008	(0.0292)	0.9713
22	CNAE6810201	0.4460	1.5621	56	CNAE4530705	(0.0332)	0.9674
23	OPCAO_SIMPLES8	0.4414	1.5550	57	CNAE4399199	(0.0333)	0.9673
24	CNAE9312300	0.4141	1.5129	58	CNAE4541205	(0.0372)	0.9635
25	NUMPARTIDO43	0.4111	1.5085	59	CNAE8532500	(0.0378)	0.9629
26	CNAE6920602	0.4096	1.5062	60	NATURJURID4120	(0.0422)	0.9587
27	CNAE8650099	0.4084	1.5044	61	QUALIFSOCIO52	(0.1089)	0.8968
28	CNAE4721102	0.3905	1.4778	62	SalBrutoF	(0.1290)	0.8790
29	CNAE8630503	0.3705	1.4484	63	InabilitadosTCDF	(0.1427)	0.8670
30	CNAE9529199	0.3662	1.4422	64	PORTEEMPRESA01	(0.2248)	0.7987
31	CNAE9430800	0.3628	1.4374	65	QTD_CNAES	(0.2293)	0.7951
32	CNAE8220200	0.3376	1.4016	66	DiasNaSoc	(0.2919)	0.7469
33	CNAE6021700	0.3286	1.3891	67	QTFCSIAPEe	(0.3633)	0.6953
34	NATURJURID2038	0.3265	1.3861	68	QTCG_SIGRHe	(1.2300)	0.2923

Para compreensão da análise dos coeficientes estimados da regressão logística, cabe informar que quanto maior o valor da potência neperiana do coeficiente do atributo, ou seja, do estimador, maior a probabilidade de ocorrência e quanto menor o valor do coeficiente menor será a probabilidade.

5.3.1. Dimensão de Corrupção

Os atributos dessa dimensão tratam sobre punibilidades relativas aos servidores ou de empresas punidas que os servidores façam parte como sócios.

O atributo CEIS (Cadastro de Empresas Inidôneas e Suspensas) foi o que apresentou o segundo maior coeficiente em relação a todos os atributos da pesquisa.

Portanto, é o atributo que melhor explica a variável dependente corrupção nessa dimensão, em consequência é o fator de risco mais expressivo nessa dimensão. A interpretação matemática desse valor é dada pelo valor numérico de potência neperiana do coeficiente desse atributo que foi de 3,0791, que pode ser interpretado como ganho de 207,9% em relação ao atributo alvo (corrupção).

O atributo CEPIM (Entidades Privadas sem Fins Lucrativos Impedidas de contratar com a Administração Pública) foi excluído por registrar variância zero, ou seja, para a presente pesquisa não houve ocorrências que pudessem indicar algum padrão relacionado à variável dependente (TARGET_CORRUPCAO).

O último atributo dessa dimensão, inabilitadosTCDF, indica a relação de pessoas inabilitadas para o exercício de cargo em comissão ou função de confiança da Administração

Pública do DF devido a irregularidades graves constatadas. Esse valor apresenta-se com pouco expressividade em relação aos 68 atributos considerados (63ª posição) e assim corresponde a fator de risco baixo nessa investigação.

O resultado em relação ao atributo CEIS guarda consonância com os estudos de Hanna [33] dos antecedentes de punibilidade registrados em relação à corrupção.

5.3.2. Dimensão Funcional

Quanto a dimensão funcional, oito atributos numéricos foram examinados.

O atributo que indicava o salário líquido dos servidores (SaliqF) foi excluído devido à alta correlação com salário bruto (salBrutoF) e o atributo de salário bruto como resultado da regressão logística apresentou pouca representatividade, 62ª posição entre os 68 atributos.

Os atributos QTCG_SIGRHSIAPE e QTFCSIGRHSIAPE também foram excluídos devido à alta correlação com os demais atributos dessa dimensão.

Dois atributos dessa dimensão apresentaram valores expressivos de coeficiente QTCG_SIGRH (1º) QTFCSIAPE (5º) e com pouco expressividade QTCG_SIAPE (67º) e QTFCSIGRH (68º). Os atributos com a sigla SIGRH indicam dados do sistema de recursos humanos do Governo do Distrito Federal e os atributos SIAPE do sistema de recursos humanos do Governo Federal, responsável pelo pagamento dos servidores do GDF que atuam na Segurança Pública.

Esses quatro atributos foram construídos para representar a quantidade de cargos e funções que foram atribuídas a um servidor ao longo do tempo. Desse modo, as mudanças de cargos e funções representam fator de risco considerável de forma diferenciada entre os dois sistemas de pagamento destinado aos servidores do GDF.

A avaliação de Gans-Morse [10] indica como fator de corrupção servidores com salários abaixo do mínimo básico ou em casos de cortes salariais em consequência de medidas austeras, ou crises do Estado. Para o presente caso não se verificou essa relação. O atributo salário bruto apresentou perda em relação à variável alvo de 12,1%. Talvez a não aderência desse atributo ao estudo apresentado possa ocorrer devido aos salários do GDF não representar situações abaixo do mínimo essencial.

5.3.3. Dimensão Política

Para dimensão política, eram inicialmente seis atributos categóricos, que para o processo de mineração foram convertidos em atributos binários representados por suas categorias, totalizando 69 atributos.

Após as exclusões de atributos com alta correlação e variância quase zero, restaram 32 que serão comentados em grupos a seguir.

Dos atributos com coeficientes mais expressivos registram-se os candidatos servidores com estado civil casado (14ª), com grau de instrução de nível médio completo (15ª), de partido político nº 43 (25ª) e candidatura para deputado distrital (46ª).

Os demais atributos não apresentaram valores significativos de coeficiente que se possa considerar como fator de risco quando considerados em conjunto em relação às seis variáveis iniciais que as geraram.

Os atributos dessa dimensão apresentam coeficientes com ganhos razoáveis (70,5%, 70,2%, 50,8% e 22,3%) em relação ao atributo alvo, o que mostra adequação às conclusões de Carvalho & outros [52].

5.3.4. Dimensão de Vínculos Societários

Na dimensão de vínculos societários foram dez atributos iniciais, sendo três atributos numéricos e sete categóricos, quando transformados em atributos binários, segundo suas categorias, formaram um total de 1.027 atributos nessa dimensão. Após exclusão de atributos com alta correlação, variância quase zero e de coeficientes com *p-value* maior que 0,05 restaram 54 atributos.

Os resultados da investigação indicaram que o percentual de capital social de sócios servidores é o atributo que melhor explica a variável dependente corrupção nessa dimensão, portanto, o fator de risco mais expressivo em relação aos demais atributos. O valor numérico de potência neperiana do coeficiente desse atributo foi de 1,86, que pode ser interpretado como ganho de 86% em relação ao atributo alvo (corrupção).

Os atributos relacionados ao porte da empresa indicam como menor fator de risco para microempresa em relação ao atributo alvo, 20% de perda (PORTEEMPRESA01), e em relação a outras denominações de empresas não classificadas como microempresa ou empresa de pequeno porte um ganho de 18% em relação ao atributo alvo (PORTEEMPRESA05). Desse

modo, enquanto a microempresa apresenta menor fator de risco do que as classificadas como não microempresa ou empresa de pequeno porte.

CNAE são as atividades empresariais cadastradas como áreas de atuação. O atributo que representa a quantidade de atividades (QTD_CNAES) cadastradas para determinada empresa foi considerada como baixo fator de risco, perda em relação ao atributo alvo de 20%, ou seja, a redução de uma unidade corresponderia a redução de 0,2 desse atributo.

No entanto, as atividades cadastradas como atividade principal das empresas (CNAE) representam a maior quantidade de atributos do resultado da pesquisa, 39 no total. Estes atributos registraram variações do exponencial neperiano dos coeficientes entre 2,29 e 0,96, que corresponde ao acréscimo de 129% ou redução de 4% em relação ao atributo alvo. Desse modo, as atividades CNAE listadas representam em sua maioria ganho em relação ao atributo alvo.

O atributo que representa a quantidade tempo em que o servidor faz parte na sociedade empresarial (DiasnaSoc) apresentou como resultado perda de 25% em relação ao atributo alvo o que sugere menor fator de risco.

A qualificação do sócio da empresa apresentou algumas categorias que transformadas em atributos apresentaram ganho em relação ao atributo resposta como Secretário 64% (QUALIFSOCIO18), Sócio-Administrador 59% (QUALIFSOCIO49), Sócio 27% (QUALIFSOCIO22), Titular Pessoa Física Residente ou Domiciliado no Brasil 10% (QUALIFSOCIO65) e redução na classificação de Sócio com Capital de 10% (QUALIFSOCIO52).

Os demais atributos dessa dimensão não apresentam representatividade em relação a variáveis originais (categóricas) que a geraram para avaliação de fator de risco em relação ao atributo alvo de forma que possam trazer conclusões de resultado de risco relevante para o entendimento de negócio da pesquisa.

5.4. Validar os resultados com os especialistas do TCDF

Os resultados obtidos nesta investigação mostram-se em grande parte em concordância com a literatura pesquisada e confirmada com a opinião de especialistas do TCDF.

Quanto à dimensão de corrupção, o resultado em relação ao atributo CEIS guarda consonância com os estudos de Hanna [33] dos antecedentes de punibilidade registrados em relação à corrupção.

A avaliação de Gans-Morse [10] indica como fator de corrupção servidores com salários abaixo do mínimo básico ou em casos de cortes salariais em consequência de medidas austeras, ou crises do Estado. Para o presente caso não se verificou essa relação na dimensão funcional. O atributo salário bruto apresentou perda em relação à variável alvo de 12,1%. Talvez a não aderência desse atributo ao estudo apresentado possa ocorrer devido aos salários do GDF não representarem situações abaixo do mínimo essencial.

Segundo especialistas, isso pode estar relacionado ao fato dos salários dos servidores do GDF não apresentar padrões de níveis próximos aos estabelecidos no estudo. O salário bruto médio dos servidores do GDF é de R$ 9.473,91, o primeiro quartil é de R$ 4.651,28 e o salário-mínimo atual (nov./2020) é de R$ 1.045,00.

O estudo de Carvalho [48] indica como fator de risco de corrupção a participação acionária de servidores em empresas, o resultado obtido nessa investigação está alinhado com esse estudo e sugere o percentual de capital social do sócio servidor do GDF como fator de risco de corrupção relevante (8ª posição com ganho de 86,9% em relação ao atributo alvo) e com maior grau o servidor sócio de empresa que consta na lista do Cadastro de Empresas Inidôneas e Suspensas – CEIS (2ª posição com ganho de 207,9% em relação ao atributo alvo).

Para dimensão política constatou-se que quatro atributos representam risco em relação ao atributo alvo. A relação de corrupção com a filiação partidária também é informada pelo estudo de Bersch & Matthew [46] que identificam que o domínio partidário em nomeações políticas está associado à menor capacidade de agência, também acrescentam que o controle político afeta de forma negativa os incentivos da carreira burocrata.

No mesmo sentido, Meyer-Sahling & Mikkelsen [47] identificam que quanto maior a politização, indicação política de cargos, maior é o nível de corrupção; enquanto quanto maior o mérito do recrutamento, menor é a corrupção associada.

Os especialistas consultados entendem o resultado como válido e após discussão concluiu-se que o cenário de maior risco representaria os servidores com as seguintes características:

- ✓ Sócios de empresas presentes na lista de Cadastro de Empresas Inidôneas e Suspensas (CEIS);
- ✓ Diversas mudanças de cargos na área de Segurança Pública ou funções públicas fora da área de Segurança Pública;
- ✓ Sócios de empresas com maior percentual acionário;

- ✓ Socio de empresas em algumas atividades específicas como: instituição de educação superior, laboratórios clínicos, serviços advocatícios, comércio de móveis e suvenires;
- ✓ Participação a cargo eletivo.

Com menor risco de corrupção elencou-se as seguintes características:

- ✓ Nunca foi penalizado;
- ✓ Não é sócio de empresa;
- ✓ Apenas um cargo ou função pública no período de vida laboral;
- ✓ Não ter participado de cargo eletivo.

Os especialistas do TCDF ratificaram o resultado atestando a validade do modelo e considerando que os coeficientes da regressão logística podem e devem ser utilizados para elaboração de planejamento de fiscalizações considerando-os como fatores de risco na priorização de ações de controle externo e, ainda, que podem trazer benefícios a instituição.

6. Conclusão e Trabalhos futuros

Nesse trabalho foi realizado o estudo e a aplicação de técnicas de mineração de dados para criação de modelo preditivo para avaliação de risco de corrupção de servidores públicos do Distrito Federal com intuito de obter os fatores de risco para subsidiar as ações de fiscalização do TCDF no âmbito de sua competência.

Conforme objetivo específico 1, elaborou-se revisão de literatura que permitiu identificar os fatores de risco de corrupção de servidores públicos avaliados em diversos países o que permitiu sugerir atributos para presente pesquisa com a opinião de especialistas (item 2.1).

A partir de outra pesquisa de literatura, definida no objetivo específico 2, levantou-se as técnicas de mineração para cenários de corrupção ou fraude (item 2.2), desse estudo e da avaliação dos dados e atributos disponíveis foi definida a aplicação de regressão logística para investigação.

O objetivo específico 3 tratou especificamente da atividade de obtenção, tratamento e carga dos dados e criação de modelos. O trato com dados foi um processo moroso que dependeu de especialistas para entendimento dos dados. Diversas bases de dados foram utilizadas para compor o conjunto de dados, a maior parte constituída de dados abertos. As ações para concretização desse objetivo foram apresentadas no capítulo 3.

Após compilação das diversas bases de dados para formação dos dados de mineração para identificação dos padrões de corrupção a partir de aprendizagem de máquina supervisionada foi reconhecida a característica aos cenários de fraude e corrupção que é o desbalanceamento de dados.

Essa dificuldade apresentada no processo de mineração de dados representou um desafio, pois os dados evidenciaram um desbalanceamento em situação extrema e, nesse cenário, obter resultados de performance adequados a geração de conhecimento dependeu de aplicação de técnicas específicas para condições extremas.

Após aplicação de técnicas de reamostragem com uso de *synthetic minority oversampling technique SMOTE* e aplicação de algoritmos com características específicas de parametrização para obter os padrões desejados da classe minoritária que corresponde a classe de interesse da pesquisa, sem que o algoritmo tendencie para classificar pela classe dominante.

Dos métodos empregados com uso de tratamento de dados de forma a equilibrar as classes, não foi possível obter desempenho adequado para modelagem com uso de SMOTE, possivelmente devido à condição de desbalanceamento extremo. Esses resultados estão descritos no capítulo 4 dessa dissertação.

Contudo, para aplicação de regressão logística com parametrização específica para o contexto de desbalanceamento extremo, obteve-se um resultado dessa pesquisa de forma adequada para modelagem permitindo gerar os coeficientes da regressão logística e interpretá-los como fatores de risco de corrupção para aplicação no planejamento de fiscalizações do TCDF e validadas pelos especialistas que concordaram também com a literatura relativa à corrupção.

No capítulo 4 foram delineados os resultados dos quatro objetivos específicos, inclusive, a validação do modelo com especialistas, consignado no objetivo específico 4.

Cabe frisar que o constante monitoramento para identificação de novos padrões mostra-se necessário para manter o modelo adequado às novas tendências. Para isso, o acompanhamento de noticiários e a troca de informações entre os diversos órgãos de fiscalização e investigação no âmbito de parcerias representam fontes importantes para continuidade da evolução do modelo com acréscimo de novas bases, novos padrões de comportamento e possível avaliação de melhores técnicas de aprendizado de máquina para aprimoramento do modelo.

Nesse sentido, o modelo apresentado não pode ser considerado como produto final e acabado, mas algo que deve ser aprimorado constantemente com novos padrões de comportamento com as bases de dados disponíveis e com novas bases a serem incorporadas.

Melhorias futuras a esta pesquisa poderão ser realizadas com remodelagem incluindo atributos do sistema de compras governamentais do GDF (Ecompras) que está em processo de implementação e previsão de operacionalização em 2021.

Outro aspecto poderá ser a inclusão do resultado do modelo nos sistemas de fiscalização como um dos fatores de risco utilizados nas fiscalizações para auditoria baseada em riscos.

Referências

[1] B. A. Olken, “Monitoring Corruption: Evidence from a Field Experiment in Indonesia,” *J. Polit. Econ.*, vol. 115, no. 2, pp. 200–249, Apr. 2007, doi: 10.1086/517935.

[2] P. Mauro, “Corruption and Growth,” *Source Q. J. Econ.*, vol. 110, no. 3, pp. 681–712, 1995.

[3] Transparency International, “Transparency International - What is Corruption?” https://www.transparency.org/what-is-corruption (accessed Jun. 16, 2019).

[4] Brasil, “Lei nº 8429, de 2 de julho de 1992,” *DOU*, 1992. http://www.planalto.gov.br/ccivil_03/leis/l8429.htm (accessed Jun. 16, 2019).

[5] MPF, “Entenda o caso: Caso Lava Jato,” *Ministério Público Federal*. 2016, [Online]. Available: http://lavajato.mpf.mp.br/entenda-o-caso.

[6] S. F. Moro, “Preventing systemic corruption in Brazil,” *Daedalus*, vol. 147, no. 3, pp. 157–168, 2018, doi: 10.1162/DAED_a_00508.

[7] A. J. A. Padula and P. H. M. Albuquerque, “Government corruption on Brazilian capital markets: A study on Lava Jato (Car Wash) investigation [Corrupção governamental no mercado de capitais: Um estudo acerca da operação Lava Jato] [Corrupción gubernamental en el mercado de capitales: Un estudio ace,” *RAE Rev. Adm. Empres.*, vol. 58, no. 4, pp. 405–417, 2018, doi: 10.1590/S0034-759020180406.

[8] C. Victor and J. Faccioni, “O Papel dos Tribunais de Contas e a Sociedade,” no. camada 4, pp. 64–71, 2015.

[9] ACFE, “Report to the Nation: Occupational Fraud and Abuse. Association of Certified Fraud Examiners (2014),” p. 31, 2014.

[10] J. Gans-Morse, M. Borges, A. Makarin, T. Mannah-Blankson, A. Nickow, and D. Zhang, “Reducing bureaucratic corruption: Interdisciplinary perspectives on what works,” *World Dev.*, vol. 105, pp. 171–188, 2018, doi: 10.1016/j.worlddev.2017.12.015.

[11] C. S. C. Branco, “History of data collection and processing for Oversight at TCU - 1995-2014,” pp. 12–21, 2014.

[12] D. Kisly, A. Tereso, and M. S. Carvalho, “Implementation of multiple criteria decision

analysis approaches in the supplier selection process: A case study," *Adv. Intell. Syst. Comput.*, vol. 444, pp. 951–960, 2016, doi: 10.1007/978-3-319-31232-3_90.

[13] T. Persons, "Data analytics and the fight against corruption," *Rev. TCU*, vol. January/ap, no. 135, pp. 8–11, 2016.

[14] Brasil, "Lei nº 9883/99." http://www.planalto.gov.br/ccivil_03/LEIS/L9883.htm (accessed Jun. 09, 2019).

[15] ABNT-NBR/ISO 31.000:2018, "NBR ISO 31000:2018 Gestao de Riscos," no. 2 ed., pp. 1–17, 2018.

[16] Câmara Legislativa do Distrito Federal, "LEI COMPLEMENTAR Nº 01 DE 9 DE MAIO DE 1994.," *1994*, 1994. http://www.tc.df.gov.br/ice4/legislacao/lc-1994-00001.html (accessed Dec. 02, 2018).

[17] "Lei Complementar 840 de 23/12/2011." http://www.sinj.df.gov.br/sinj/Norma/70196/Lei_Complementar_840_23_12_2011.html (accessed May 20, 2020).

[18] R. Balaniuk, "TCU-A Mineração de dados como Apoio ao Controle Externo-282-543-1-SM (httprevista.tcu.gov.brojsindex.phpRTCUarticleviewFile282295)," 2010.

[19] Brasília., "Lei nº 6.112, de 02 de fevereiro de 2018.," pp. 6–9, 2018.

[20] ABNT, "ABNT BNR ISO 37001 - Sistemas de gestão antissuborno - Requisitos com orientação para uso," *Associação Brasileira de Normas Técnicas*. p. 63, 2017, doi: 01.080.10; 13.220.99.

[21] A. C. Gil, *Como elaborar projetos de pesquisa*, 4th ed. São Paulo, 2002.

[22] E. L. da Silva and E. M. Menezes, *Metodologia da Pesquisa e Elaboração de Dissertação*, 4. Florianópolis: Universidade Federal de Santa Catarina - UFSC, 2005.

[23] N. Mordant, J. Delour, E. Lévêque, A. Arnéodo, and J.-F. Pinton, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002, doi: 10.1613/jair.953.

[24] J. N. Mandrekar, "Receiver operating characteristic curve in diagnostic test assessment," *J. Thorac. Oncol.*, vol. 5, no. 9, pp. 1315–1316, 2010, doi: 10.1097/JTO.0b013e3181ec173d.

[25] Usama Fayyad and Ramasamy Uthurusam y, "DFrom data mining to knowledge

discovery: an overview," *AAAI Press / MIT Press.*, vol. 39, no. 11, pp. 24–26, 1996, doi: 10.1142/9789814447331_0034.

[26] SAS Institute., "Data Mining and SEMMA :: Data Mining Using SAS(R) Enterprise Miner(TM): A Case Study Approach, Third Edition." http://support.sas.com/documentation/cdl/en/emcs/66392/HTML/default/viewer.htm#n0pejm83csbja4n1xueveo2uoujy.htm (accessed Jun. 14, 2019).

[27] P. Chapman *et al.*, "CRISP-DM 1.0 Step-by-step," *ASHA Present.*, p. 73, 2000, doi: 10.1109/ICETET.2008.239.

[28] A. Azevedo and M. F. Santos, "KDD , SEMMA AND CRISP-DM : A PARALLEL OVERVIEW Ana Azevedo and M . F . Santos," *IADIS Eur. Conf. Data Min.*, pp. 182–185, 2008, [Online]. Available: http://recipp.ipp.pt/bitstream/10400.22/136/3/KDD-CRISP-SEMMA.pdf.

[29] A. M. Mariano and M. R. Santos, "Revisão da Literatura: Apresentação de uma Abordagem Integradora," *AEDEM Int. Conf.*, no. September, pp. 427–443, 2017, [Online]. Available: https://www.researchgate.net/publication/319547360%0Ahttps://www.researchgate.net/profile/Ari_Mariano/publication/319547360_Revisao_da_Literatura_Apresentacao_de_uma_Abordagem_Integradora/links/59beb024aca272aff2dee36f/Revisao-da-Literatura-Apresentacao-d.

[30] C. J. Anderson and Y. V Tverdova, "Corruption, political allegiances, and attitudes toward government in contemporary democracies," *Am. J. Pol. Sci.*, vol. 47, no. 1, pp. 91–109, Jan. 2003, doi: 10.1111/1540-5907.00007.

[31] B.-C. Liu and T. L.-P. Tang, "Does the Love of Money Moderate the Relationship between Public Service Motivation and Job Satisfaction? The Case of Chinese Professionals in the Public Sector," *PUBLIC Adm. Rev.*, vol. 71, no. 5, pp. 718–727, 2011, doi: 10.1111/j.1540-6210.2011.02411.x.

[32] J. P. Burns and W. Xiaoqi, "Civil Service Reform in China : Impacts on Civil Servants ' Behaviour," vol. 102, no. 102, pp. 58–78, 2019, doi: 10.1017/S030574100999107X.

[33] R. Hanna and S. Y. Wang, "Dishonesty and selection into public service: Evidence from India," *Am. Econ. J. Econ. Policy*, vol. 9, no. 3, pp. 262–290, Aug. 2017, doi: 10.1257/pol.20150029.

[34] P. J. C. Lassou and T. Hopper, “Government accounting reform in an ex-French African colony: The political economy of neocolonialism,” *Crit. Perspect. Account.*, vol. 36, pp. 39–57, Apr. 2016, doi: 10.1016/j.cpa.2015.10.006.

[35] A. Asoni, “Protection of property rights and growth as political equilibria,” *J. Econ. Surv.*, vol. 22, no. 5, pp. 953–987, 2008, doi: 10.1111/j.1467-6419.2008.00554.x.

[36] T. Dodge, “State and society in Iraq ten years after regime change: the rise of a new authoritarianism,” *Int. Aff.*, vol. 89, no. 2, SI, pp. 241–257, Mar. 2013, doi: 10.1111/1468-2346.12016.

[37] O. Poocharoen and A. Brillantes, “Meritocracy in Asia Pacific: Status, Issues, and Challenges,” *Rev. PUBLIC Pers. Adm.*, vol. 33, no. 2, SI, pp. 140–163, Jun. 2013, doi: 10.1177/0734371X13484829.

[38] M. Génaux, “Social sciences and the evolving concept of corruption,” *Crime, Law and Social Change*, vol. 42, no. 1. pp. 13–24, Sep. 2004, doi: 10.1023/B:CRIS.0000041034.66031.02.

[39] B. Rubbers, “The ``informal sector{''}: The economy of Katanga (Congo-Zaire) and the falsification of the law,” *Sociol. Trav.*, vol. 49, no. 3, pp. 316–329, 2007, doi: 10.1016/j.soctra.2007.06.024.

[40] K. T. Liou, L. Xue, and K. Dong, “China’s Administration and Civil Service Reform: An Introduction,” *Rev. PUBLIC Pers. Adm.*, vol. 32, no. 2, SI, pp. 108–114, Jun. 2012, doi: 10.1177/0734371X12438241.

[41] T. Gong and J. Ren, “Hard Rules and Soft Constraints : regulating conflict of interest in China Hard Rules and Soft Constraints : regulating conflict of interest in China,” vol. 0564, 2013, doi: 10.1080/10670564.2012.716941.

[42] D. Treisman, “What have we learned about the causes of corruption from ten years of cross-national empirical research?,” 2007, doi: 10.1146/annurev.polisci.10.081205.095418.

[43] C. Van Rijckeghem and B. Weder, “Bureaucratic corruption and the rate of temptation: do wages in the civil service affect corruption, and by how much?,” 2001. [Online]. Available: www.elsevier.comrlocatereconbase.

[44] J. E. Rauch and P. B. Evans, “Bureaucratic structure and bureaucratic performance in less developed countries,” vol. 75, pp. 49–71, 2000.

[45] T. B. Pepinsky, J. H. Pierskalla, and A. Sacks, “Bureaucracy and Service Delivery,” in *ANNUAL REVIEW OF POLITICAL SCIENCE, VOL 20*, vol. 20, 4139 EL CAMINO WAY, PO BOX 10139, PALO ALTO, CA 94303-0897 USA: ANNUAL REVIEWS, 2017, pp. 249–268.

[46] K. Bersch, S. Praça, and M. M. Taylor, “State Capacity, Bureaucratic Politicization, and Corruption in the Brazilian State,” *Governance*, vol. 30, no. 1, pp. 105–124, Jan. 2017, doi: 10.1111/gove.12196.

[47] J. H. Meyer-Sahling and K. S. Mikkelsen, “CIVIL SERVICE LAWS, MERIT, POLITICIZATION, AND CORRUPTION: THE PERSPECTIVE OF PUBLIC OFFICIALS FROM FIVE EAST EUROPEAN COUNTRIES,” *Public Adm.*, vol. 94, no. 4, pp. 1105–1123, Dec. 2016, doi: 10.1111/padm.12276.

[48] R. S. Carvalho, “Modelos preditivos para avaliação de risco de corrupção de servidores públicos federais,” 2015, [Online]. Available: http://repositorio.unb.br/handle/10482/19361.

[49] R. N. Carvalho and R. S. Carvalho, “Bayesian Models to Assess Risk of Corruption of Federal Management Units,” *Proc. 13th UAI Bayesian Model. Appl. Work.*, no. 8, pp. 28–35, 2016.

[50] B. G. López-valcárcel, J. Luis, and J. Perdiguero, “Danger: Local corruption is contagious!,” *J. Policy Model.*, vol. 39, no. 5, pp. 790–808, 2017, doi: 10.1016/j.jpolmod.2017.08.002.

[51] K. H. Pedersen and L. Johannsen, “Where and How You Sit: How Civil Servants View Citizens’ Participation,” *Adm. Soc.*, vol. 48, no. 1, pp. 104–129, Jan. 2016, doi: 10.1177/0095399714555753.

[52] R. Carvalho, R. Carvalho, M. Ladeirat, F. Monteiro, and G. Mendes, “Using Political Party Affiliation Data to Measure Civil Servants’ Risk of Corruption,” in *2014 BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS (BRACIS)*, 2014, pp. 166–171, doi: 10.1109/BRACIS.2014.39.

[53] P. M. Buscema *et al.*, “The perception of corruption in health: AutoCM methods for an international comparison,” *Qual. Quant.*, vol. 51, no. 1, pp. 459–477, 2017, doi: 10.1007/s11135-016-0315-4.

[54] A. Rashidian, H. Joudaki, and T. Vian, “No evidence of the effect of the interventions

to combat health care fraud and abuse: A systematic review of literature," *PLoS One*, vol. 7, no. 8, 2012, doi: 10.1371/journal.pone.0041988.

[55] Z. Chen, L. D. Van Khoa, A. Nazir, E. N. Teoh, and E. K. Karupiah, "Exploration of the effectiveness of expectation maximization algorithm for suspicious transaction detection in anti-money laundering," *ICOS 2014 - 2014 IEEE Conf. Open Syst.*, pp. 145–149, 2014, doi: 10.1109/ICOS.2014.7042645.

[56] J. Tang, "On developing intelligent surveillant system of suspicious financial transaction," *2010 2nd Int. Conf. E-bus. Inf. Syst. Secur. EBISS2010*, pp. 592–595, 2010, doi: 10.1109/EBISS.2010.5473748.

[57] A. A. Rizki, I. Surjandari, and R. A. Wayasti, "Data mining application to detect financial fraud in Indonesia's public companies," *Proceeding - 2017 3rd Int. Conf. Sci. Inf. Technol. Theory Appl. IT Educ. Ind. Soc. Big Data Era, ICSITech 2017*, vol. 2018-Janua, pp. 206–211, 2018, doi: 10.1109/ICSITech.2017.8257111.

[58] C. Ghedini Ralha and C. V. Sarmento Silva, "A multi-agent data mining system for cartel detection in Brazilian government procurement," *Expert Syst. Appl.*, vol. 39, no. 14, pp. 11642–11656, 2012, doi: 10.1016/j.eswa.2012.04.037.

[59] R. Balaniuk, P. Bessiere, E. Mazer, and P. R. Cobbe, "Corruption risk analysis using semi-supervised naïve Bayes classifiers," *Int. J. Reason. Intell. Syst.*, vol. 5, no. 4, p. 237, 2014, doi: 10.1504/ijris.2013.058768.

[60] S. L. Domingos, R. N. Carvalho, R. S. Carvalho, and G. N. Ramos, "Identifying it purchases anomalies in the Brazilian Government Procurement System using deep learning," *Proc. - 2016 15th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2016*, no. Cic, pp. 722–727, 2017, doi: 10.1109/ICMLA.2016.106.

[61] V. Shehu, A. Mijushkovic, and A. Besimi, "Empowering Data Driven Journalism in Macedonia," no. 335, pp. 1–6, 2016, doi: 10.1145/2955129.2955187.

[62] J. Q. Su and S. Dan, "Application of Data Mining in Construction of Corruption Risks Prevention System," *Appl. Mech. Mater.*, vol. 513–517, pp. 2165–2169, 2014, doi: 10.4028/www.scientific.net/amm.513-517.2165.

[63] J. Su and S. Dan, "Research on framework of corruption risks prevention system based on cloud computing," *2013 2nd Int. Symp. Instrum. Meas. Sens. Netw. Autom.*, pp. 141–144, 2014, doi: 10.1109/imsna.2013.6743236.

[64] H. A. A. Arief, G. A. P. Saptawati, and Y. D. W. Asnar, “Fraud detection based-on data mining on Indonesian E-Procurement System (SPSE),” *Proc. 2016 Int. Conf. Data Softw. Eng. ICoDSE 2016*, pp. 1–6, 2017, doi: 10.1109/ICODSE.2016.7936111.

[65] M. T. Islam and M. A. Yousuf, “Development of a Corruption Detection Algorithm using K-means Clustering,” *2018 Int. Conf. Adv. Electr. Electron. Eng. ICAEEE 2018*, pp. 1–4, 2019, doi: 10.1109/ICAEEE.2018.8642985.

[66] Y. Rong, E. Xu, and N. Li, “Mining a microblog network on anti-corruption news with Social Network Analysis,” *2015 12th Int. Conf. Fuzzy Syst. Knowl. Discov. FSKD 2015*, pp. 1432–1436, 2016, doi: 10.1109/FSKD.2015.7382154.

[67] X. Qin, Z. Jin, and F. Fangyu, “Research on Anti-Corruption Public Opinion Data Analysis Technology Based on SPSS under ‘Internet +’ Environment,” pp. 59–64, 2018, doi: 10.1145/3210506.3210517.

[68] J. Friedman, T. Hastie, and R. Tibshirani, *The elements os Statistical Learning: Data Mining, Inference and Prediction*. 2001.

[69] C. C. Aggarwal, *Data Mining*. New York, 2015.

[70] B. Zhu, B. Baesens, A. Backiel, and S. K. L. M. Vanden Broucke, “Benchmarking sampling techniques for imbalance learning in churn prediction,” *J. Oper. Res. Soc.*, vol. 69, no. 1, pp. 49–65, 2018, doi: 10.1057/s41274-016-0176-1.

[71] J. Brownlee, “Imbalanced Classification with Python Choose Better Metrics, Balance Skewed Classes, and Apply Cost-Sensitive Learning,” *Mach. Learn. Mastery*, pp. 1–22, 2020.

[72] A. Melo Mariano and M. Rocha Santos, *Revisão da Literatura: Apresentação de uma Abordagem Integradora Structural Equations View project Service Quality View project*. 2017.

[73] P. Branco, L. Torgo, and R. P. Ribeiro, “A survey of predictive modeling on imbalanced domains,” *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1–50, 2016, doi: 10.1145/2907070.

[74] M. Buda, A. Maki, and M. A. Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks,” *Neural Networks*, vol. 106, pp. 249–259, 2018, doi: 10.1016/j.neunet.2018.07.011.

[75] Alberto Fernandez, Salvador Garcia, Francisco Herrera, and Nitesh V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018, [Online]. Available: https://www.jair.org/index.php/jair/article/view/11192.

[76] W. Mao, L. He, Y. Yan, and J. Wang, "Online sequential prediction of bearings imbalanced fault diagnosis by extreme learning machine," *Mech. Syst. Signal Process.*, vol. 83, pp. 450–473, 2017, doi: 10.1016/j.ymssp.2016.06.024.

[77] J. Sun, J. Lang, H. Fujita, and H. Li, "Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates," *Inf. Sci. (Ny).*, vol. 425, pp. 76–91, 2018, doi: 10.1016/j.ins.2017.10.017.

[78] J. A. Sáez, B. Krawczyk, and M. Woźniak, "Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets," *Pattern Recognit.*, vol. 57, pp. 164–178, 2016, doi: 10.1016/j.patcog.2016.03.012.

[79] Z. Feng *et al.*, "Machine learning-based quantitative texture analysis of CT images of small renal masses: Differentiation of angiomyolipoma without visible fat from renal cell carcinoma," *Eur. Radiol.*, vol. 28, no. 4, pp. 1625–1633, 2018, doi: 10.1007/s00330-017-5118-z.

[80] S. Guo, R. Chen, H. Li, T. Zhang, and Y. Liu, "Identify Severity Bug Report with Distribution Imbalance by CR-SMOTE and ELM," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 29, no. 2, pp. 139–175, 2019, doi: 10.1142/S0218194019500074.

[81] T. Package, T. Functions, and A. L. Torgo, *Package ' DMwR . '* 2015.

[82] L. Torgo, *Data Mining with R - Learning with Case Studies*. Minneapolis, Minnesota: Data Mining and Knowledge Discovery Series.

[83] D. Hosmer and S. Lemeshow, "Applied Survival Analysis - Regression Modeling of Time to Event Data." 1999.

[84] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python and Scikit-Learn*. 2017.