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

Modelo Preditivo de Risco de Irregularidades em Compras Públicas no Estado de Goiás

Mauricio Barros de Jesus

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

Orientador
Prof. Dr. Gladston Luiz da Silva

Brasília
2020

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

Barros de Jesus, Mauricio
BD278m Modelo Preditivo de Risco de Irregularidades em Compras Públicas no Estado de Goiás / Mauricio Barros de Jesus; orientador Gladston Luiz da Silva. -- Brasília, 2020.
105 p.

Dissertação (Mestrado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2020.

1. mineração de dados. 2. irregularidades. 3. licitações. 4. seleção de variáveis. I. Luiz da Silva, Gladston, orient. II. Título.

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

Modelo Preditivo de Risco de Irregularidades em Compras Públicas no Estado de Goiás

Mauricio Barros de Jesus

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

Prof. Dr. Gladston Luiz da Silva (Orientador)
Diretor do Instituto de Ciências Exatas/UnB

Prof. Dr. Donald Matthew Pianto
IE/UnB

Prof. Dr. Sávio Salvarino Teles de Oliveira
IPOG

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

Brasília, 17 de dezembro de 2020

Dedicatória

Ao Sr. João Pedro de Barros, que me transmitiu sabiamente os valores de fé, família, honestidade e perseverança.

Agradecimentos

Agradeço a Deus por me conceder saúde e condições para enfrentar o desafio de cursar um programa de pós-graduação.

Aos auditores Sérvio Túlio, Bruno Henrique, Luzia Dourado e Rodrigo Cruvinel pela torcida e por dedicar tempo em discussões técnicas e revisões de dados.

À Ms. Raquel Luz pelo incentivo, companheirismo e apoio incondicional, mesmo nos momentos mais difíceis.

Ao meu amigo Vitor Gobato, por acreditar no meu trabalho e por prestigiar o aperfeiçoamento dos seus subordinados.

Ao Prof. Dr. Gladston, por me receber como orientando e pelas contribuições dadas na pesquisa de forma objetiva e paciente.

Aos professores do Programa de Pós-graduação em Computação Aplicada (PPCA), pelo conhecimento repassado, e aos meus colegas de mestrado, pelas discussões e contribuições nessa jornada que compartilhamos.

Ao Tribunal de Contas do Estado de Goiás (TCE-GO), instituição em trabalho com orgulho e dedicação, por todo apoio logístico ofertado.

Resumo

O exercício eficaz do controle externo, dada as limitações de recursos públicos, deve acontecer em um cenário em que o esforço de fiscalização se concentra nos casos de maior probabilidade de risco de se encontrar irregularidades. No tocante às compras governamentais, na ausência de uma métrica objetiva de seleção de licitações para fins de auditoria, este trabalho apresenta uma proposta de modelo preditivo para estabelecer um indicador de risco de irregularidades em compras públicas em dois momentos da licitação: publicação do edital e disputa. A partir desse indicador foi construído um *ranking* de licitações com indícios de irregularidades que pode ser utilizado como ferramenta para tomada de decisão nas auditorias. Para isso, foram utilizadas licitações realizadas nos Estado de Goiás nos anos de 2014 à 2019, coletados em bases de dados que o TCE-GO tem acesso. Inicialmente foi realizado um estudo da fundamentação legal das irregularidades em compras públicas com vistas a identificar quais atributos que podem influenciar no risco do certame e suas respectivas bases de dados. A partir dos dados levantados, foram aplicados quatro métodos de seleção de atributos para identificar as variáveis mais importantes e em que medida contribuem para explicar o risco da licitação. O risco foi calculado utilizando técnicas de treinamento supervisionado com quatro classificadores. Como resultado, foi constatado que modelos especialistas por modalidade de licitação têm melhor desempenho do que modelos genéricos treinados com todo o conjunto de dados. Para a fase de publicação do edital, licitações da modalidade pregão, dispensa e inexigibilidade apresentaram resultados de AUROC acima de 70%, no entanto, a modalidade concorrência não apresentou resultados aceitáveis. Para a etapa de disputa, todas as modalidades tiveram AUROC acima de 70%.

Palavras-chave: mineração de dados, irregularidades, licitações, seleção de variáveis.

Abstract

The effective exercise of external control must be optimized so that the inspection effort is concentrated in cases of high risk of irregularities. With regard to government procurement, in the absence of an objective metric for selecting public procurement for audit purposes, this paper presents a proposal for a predictive model to establish an indicator of the risk of irregularities in public procurement in two phases: "Publication" phase and "Dispute" phase. A ranking of public purchases classified by the probability of irregularities was built. This ranking can be used as a tool for decision making in audits. Public purchases made by the State of Goiás between 2014 and 2019, which were collected in databases to which the TCE-GO has access. Initially, a study was carried out on the legal requirements of irregularities in public purchases, in order to identify which attributes may influence the risk of irregularity and which are the most important databases Four methods of attribute selection were applied to identify the variables Most important and to identify how much these variables contribute to explain the risk in public purchases. The risk was calculated using supervised training techniques with four classifiers. As a result, it was found that the specialized models by bidding modality performed better than the generic models trained with the entire data set. For the "Publication" phase, the "Pregão", "Dispensa" and "Inexigibilidade" modalities presented AUROC results above 70%, but the "Concorrência" modality did not present acceptable results. For the "Dispute" phase, all modalities had AUROC above 70%..

Keywords: data mining, corruption, public purchase, feature selection

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AUROC *Area Under the Receiver Operating Characteristic Curve*.

CACC *Class-Attribute Contingency Class*.

Ceasa-GO Centrais de Abastecimento de Goiás.

CEIS Cadastro Nacional de Empresas Inidôneas e Suspensas.

CEPIM Cadastro de Entidades Privadas Sem Fins Lucrativos Impedidas.

CGU Controladoria Geral da União.

CNAE Classificação Nacional de Atividades Econômicas.

CNEP Cadastro Nacional de Empresas Punidas.

Compras-NET Base de dados de licitações do Poder Executivo do Estado de Goiás.

CRISP-DM *Cross-Industry Standard Process for Data Mining*.

DAV Conjunto de dados para avaliação.

DTT Conjunto de para treino/teste dos modelos.

e-TCE Sistema de Controle de Processos do Tribunal de Contas do Estado de Goiás.

ElasticNetCV *Elastic-Net* com *Cross-Validation*.

EPP Empresa de Pequeno Porte.

ETL Extração, Transformação e Carga.

GBM *Gradient Boosting* para classificação.

Goinfra Agência Goiana de Infraestrutura e Transportes.

Iquego Indústria Química do Estado de Goiás.

JUCEG Junta Comercial do Estado de Goiás.

Lars *Least Angle Regression*.

LarsCV *Least Angle Regression* com *Cross-Validation*.

LASSO *Least Absolute Shrinkage and Selection Operator*.

LASSOCV *Least Absolute Shrinkage and Selection Operator* com *Cross-Validation*.

MDLP *Minimum Description Length Principle*.

ME Microempresa.

MLG Modelos Lineares Generalizados.

MSE Erro Quadrático Médio.

PCA *Principal Component Análise*.

PDCA *Plan, Do, Check, Action*.

RAIS Relação Anual de Informações Sociais.

RFE *Recursive Feature Elimination*.

RH-NET Base de dados cadastral dos servidores do Poder Executivo do Estado de Goiás.

ROC *Receiver Operating Characteristic*.

Saneago Companhia Saneamento de Goiás S/A.

SGF Sistema de Gestão de Fiscalização do TCE-GO.

SIOFI-NET Sistema de Execução Financeira e Orçamentária do Estado de Goiás.

SMOTE *Synthetic Sampling with Data Gereration*.

STF Supremo Tribunal Federal.

SVM *Support Vector Machine*.

TCE-GO Tribunal de Contas do Estado de Goiás.

TCU Tribunal de Contas da União.

TSE Tribunal Superior Eleitoral.

Capítulo 1

Introdução

A fraude e a corrupção consomem recursos que deveriam ser aplicados em benefício da sociedade. Tratam-se de fenômenos presentes em todos os países, mas mesmo assim não existe um conceito único para esses termos, pois cada país os definem de acordo com fatores sociais, éticos e morais [1]. No Brasil, o combate à fraude e à corrupção são temas relevantes na agenda governamental. A partir da promulgação da Carta Magna de 1988 [1], o poder público editou um conjunto de normas consolidando o que ficou conhecido como Sistema Brasileiro de Combate à Corrupção [2]. Isso fez com que o Brasil fosse reconhecido internacionalmente como um país com arcabouço adequado de leis que punem o suborno, a lavagem de dinheiro, formação de carteis e o enriquecimento ilícito [3].

Em 1992 foi editada a Lei 8.429 [2], que estabeleceu critérios para punição de agentes públicos por enriquecimento ilícito de mandato, cargo, emprego ou função. Posteriormente, em 2001, o Governo Federal criou a Controladoria Geral da União (CGU), órgão de controle interno com a missão de combate à fraude e à corrupção no âmbito do Poder Executivo [3].

A CGU serviu de modelo para que os estados e municípios criassem órgãos semelhantes em suas estruturas, estabelecendo-se respectivamente as Controladorias Estaduais e as Controladorias Municipais. Dessa forma, gradativamente foi criada uma rede de prevenção e combate a corrupção no âmbito do Poder Executivo com atuação em todo território brasileiro.

Outra iniciativa foi a publicação da lei anticorrupção com aplicabilidade em todo território nacional [4], que definiu critérios para responsabilizar todos aqueles que praticam atos que possam lesar a administração pública, sejam eles pessoas físicas ou jurídicas, entida-

[1]Constituição Federal do Brasil - http://www.planalto.gov.br/ccivil _03/constituicao/ constituicao.htm

[2]Lei 8.429/1992 - http://www.planalto.gov.br/ccivil_03/leis/L8429.htm

[3]Lei 13.844/2019 - http://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2019/Lei/L13844.htm

[4]Lei 12.846/2013 - http://www.planalto.gov.br/ ccivil_03/_ato2011-2014/ 2013/lei/l12846.htm

des públicas ou privadas. Dentre os avanços alcançados está a normatização dos acordos de leniência, que deram competência à CGU para instaurar processos administrativos de responsabilização em casos de corrupção.

Nesse mesmo ano foi normatizado o instituto da colaboração premiada [5] como meio de obtenção de provas. Dessa forma, caso alguém que tenha praticado ato de corrupção, mas que de livre e espontânea vontade denuncie as pessoas envolvidas, poderá obter uma redução de pena caso a informação dada possibilite o desmantelamento do esquema fraudulento.

Nesta esteira, o Sistema Brasileiro de Combate à Corrupção e o fortalecimento das instituições de combate a fraude e corrupção foi crucial para o sucesso da Operação Lava Jato, iniciada em 2014. Trata-se de atuação conjunta do Ministério Público Federal, Poder Judiciário, órgãos de controle interno e controle externo, Receita Federal e Polícia Federal, que trouxe às claras um esquema de corrupção que, até o fim de 2018, girava em torno de R$ 39 bilhões de reais. [4].

A atuação conjunta dessas instituições públicas é necessária, pois a fraude pode envolver esquemas sofisticados e organizados para sua ocultação [4]. Por exemplo, segundo o Ministério Público Federal, na Operação Lava Jato foi identificado um esquema de pagamento de propinas que giravam entre 1% a 5% do valor dos contratos entre empresas privadas e o poder público. Acordos fraudulentos eram feitos entre grandes empreiteiras, doleiros, operadores financeiros, agentes públicos e pessoas com cargos políticos na fases iniciais dos procedimentos licitatórios para aquisição de bens se serviços [4].

Em países em desenvolvimento como o Brasil, os processos de aquisições públicas de bens e serviços estão dentre aqueles com maior risco de fraude e corrupção [5]. Além disso, a insuficiência ou a ausência de controle em áreas de risco, tanto por parte do Poder Legislativo (controle externo) quanto do Poder Executivo (controle interno), contribuem como causa raiz das irregularidades [5].

No Brasil, o Poder Legislativo é o titular do controle externo. Na esfera federal o Tribunal de Contas da União (TCU) é órgão técnico que auxilia o Congresso Nacional no exercício do controle externo. Segundo a Constituição Federal [6], os tribunais de contas são responsáveis por julgar contas de gestores e apreciar contas de chefe do Pode Executivo; realizar auditorias de aspecto contábil, orçamentário, financeiro ou patrimonial; realizar o controle de legalidade de atos de admissão de pessoal e de processos licitatórios. Sua jurisdição alcança os Poderes Legislativo, Judiciário e Executivo [7]. Todo o cenário passível de atuação dos tribunais de contas denomina-se universo de controle.

A estrutura federal é replicada nos Estados. O controle externo no Estado de Goiás é exercido pela Assembleia Legislativa com apoio do Tribunal de Contas do Estado de

[5]Lei 12.850/2013 - http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2013/Lei/L12850.htm

Goiás (TCE-GO), conforme disposto no artigo 25 da Constituição Estadual de Goiás [8]. Dentre as diversas atribuições do TCE-GO estão o exercício do controle prévio e do controle concomitante, atuando na prevenção de falhas e na adoção de medidas saneadoras antes da concretização de ato ilegal ou de dano ao erário [9].

Especificamente no que tange às compras públicas, o controle prévio e concomitante se manifesta quando o Tribunal solicita aos órgãos sob sua jurisdição documentos dos processos licitatórios em curso. Essa previsão se encontra no parágrafo 4º do art. 263 do Regimento Interno do TCE-GO, que dispõe: "os editais de licitação e os atos de dispensa e inexigibilidade serão acompanhados de forma seletiva e concomitante". Ao propor a seleção dos certames, o texto regimental respalda a necessidade de racionalização de recursos, de forma a aumentar da eficiência da máquina pública.

Por outro lado, a quantidade de aquisições feitas pelo poder público tem aumentado nos últimos anos e por consequência, a quantidade de informações que devem ser analisadas atinge volumes cada vez maiores. Em valores financeiros, no ano de 2017, no Estado de Goiás foram gastos R$ 5,2 bilhões em aquisições apenas no Poder Executivo [6].

Não obstante, dado o tamanho do universo de controle, os tribunais de contas tem se atentado para modernização dos processos de fiscalização, com uso de softwares especialistas para organização de dados e automatização de tarefas [7]. Mas mesmo com o aperfeiçoamento tecnológico, que permite maior tempestividade para detecção de indícios de irregularidades, ainda é necessária a análise manual por parte do auditor que, muitas vezes, solicita esclarecimentos ao jurisdicionado ou em outros casos, realiza fiscalização *in loco*.

Desta forma, para conferir maior efetividade às fiscalizações, agir de forma seletiva requer análise minuciosa do objeto de controle, para que o esforço de fiscalização aconteça em cenários com maior risco de encontrar irregularidades. Mas antes, é necessário um estudo profundo de quais são os fatores que podem indicar uma tentativa ou perpetuação de fraude e corrupção. Além desses fatores existe o fenômeno da ineficiência da administração púbica, caracterizada pela má gestão, por imperícia e imprudência, que, mesmo não sendo caracterizado como com fraude, geram gastos desnecessários para o poder público [10]. Todos esses casos configuram indícios de irregularidades.

Diante desse contexto, este trabalho tem como objetivo construir, utilizando dados do TCE-GO, um modelo preditivo para estabelecer um indicador de risco de irregularidades em licitações para duas fases do procedimento licitatório: publicação do edital e disputa.

A criação da solução proposta foi baseada em mineração de dados e aprendizado de máquina. Foi construída de acordo com a execução das fases modelo de referência *Cross-Industry Standard Process for Data Mining* (CRISP-DM) [11], que decompõe as

[6] http://www.transparencia.go.gov.br/portaldatransparencia/gastos-governamentais/licitacoes.

atividades de mineração de dados em 6 etapas, numa abordagem *top-down* que se inicia no entendimento do negócio e se estende até a implantação do produto [12].

Foram utilizados os dados que o TCE-GO já possui acesso: Sistema Informa - base de dados de licitações de todos os jurisdicionados; Sistema de Controle de Processos do Tribunal de Contas do Estado de Goiás (e-TCE); Base de dados de licitações do Poder Executivo do Estado de Goiás (Compras-NET); Base de dados cadastral dos servidores do Poder Executivo do Estado de Goiás (RH-NET); CPF/CNPJ da Receita Federal do Brasil - base cadastral de pessoas físicas e jurídicas; Relação Anual de Informações Sociais (RAIS); Doadores de Campanhas Eleitorais mantida pelo Tribunal Superior Eleitoral (TSE);

Complementarmente, foram utilizadas as bases de dados abertas mantidas pela CGU: Cadastro Nacional de Empresas Inidôneas e Suspensas (CEIS), que contém informações sobre as empresas inidôneas; Cadastro Nacional de Empresas Punidas (CNEP), que apresenta informações sobre as empresas punidas.

1.1 Justificativa

A otimização de recursos é vital para o controle externo no Brasil, pois o tamanho do universo de controle muitas vezes extrapola a capacidade de fiscalização disponível. Dessa forma, exigir que tudo seja alvo análise pelo Controlador cria um ambiente burocratizado, moroso, caro e com gargalos na Administração Pública. Além disso, com base no princípio constitucional da eficiência, o custo de se auditar deve ser menor do que o ganho potencial da auditoria.

No que se refere aos atos de aquisição de bens e serviços pelo poder público, tema desse trabalho, a título de exemplo, em 2017 o governo federal realizou cerca de 136 mil licitações, com um volume financeiro de aproximadamente R$ 171 bilhões de reais. Em 2018 esse número saltou para 152 mil certames [7].

Por sua vez, no Estado de Goiás, entre janeiro 2017 a dezembro de 2018 foram realizados 6.733 certames apenas pelo Poder Executivo [8], totalizando R$10 bilhões de reais. Ressalta-se que a Lei de Responsabilidade Fiscal impõe restrições aos gastos públicos em anos eleitorais e, em 2018, ocorreram as eleições estaduais. Dessa forma, tem-se que fora desse cenário esses números normalmente são maiores.

Independentemente da modalidade da licitação, a ação de controle tem maior efetividade antes da adjudicação do objeto da licitação, pois nessa etapa ainda não existe uma obrigação objetiva entre o poder público e determinada empresa vencedora [9]. Quanto

[7]Fonte: http://www.portaltransparencia.gov.br/licitacoes. Consultado em 25/10/2020
[8]Fonte: Sistema Informa do TCE-GO em https://informa.tce.go.gov.br

mais tempo se leva para reparar um potencial problema ou indício de irregularidade em uma licitação, mais cara e complexa a ação de controle se torna [9].

Logo, dada a quantidade de licitações que são realizadas frente ao curto espaço temporal para fiscalização e aos limitados recursos físicos e humanos disponíveis, o censo é inviável e antieconômico. Portanto, resta claro que existe uma lacuna de melhoria na forma como os processos licitatórios são selecionados.

Este trabalho visa preencher essa lacuna, possibilitando a priorização das fiscalizações para direcionar o esforço naqueles casos com maior potencial de se encontrar indícios de irregularidades. Além disso, os atributos selecionados nas etapas de modelagem podem contribuir com a simplificação de exigências de informação declaratórias nos sistemas do TCE-GO.

1.2 Objetivos

O objetivo geral deste trabalho é construir, utilizando dados do TCE-GO, um modelo preditivo para estabelecer um indicador de risco de irregularidades em licitações para duas fases do procedimento licitatório: publicação do edital e disputa.

Para isto, necessita-se que outros objetivos específicos sejam atendidos, a saber:

- Identificar a fundamentação legal relacionada a irregularidades em compras públicas;
- Identificar os atributos que podem influenciar no risco do certame e suas respectivas bases de dados;
- Identificar os atributos mais relevantes que caracterizam indícios irregularidades em licitações;
- Desenvolver um indicador de risco de irregularidades dos certames na fase de publicação do edital e na fase disputa, através de um modelo preditivo baseado em mineração de dados;
- Avaliar o indicador de risco junto aos especialistas de fiscalização; e
- Propor um processo automatizado para obtenção e apresentação do indicador de risco na forma de um *ranking*.

1.3 Metodologia

A partir da definição do problema de pesquisa foram estabelecidos os objetivos gerais e específicos. Foi realizada a revisão da literatura por meio de pesquisas nas bases *Web of*

Science e *Scopus*, buscando artigos publicados a partir de 2016 relacionados a *machine learning* aplicadas para fiscalização e detecção de fraudes.

Para tanto, foram utilizadas combinações de pesquisa por palavras chaves, tais como "fraud detection", "machine learning", "public purchase", "bidding", "corruption", "data mining" e "artificial intelligence". Como resultado, foram selecionados artigos com base na leitura do *abstract* e da conclusão, buscando maior aderência possível aos objetivos desta pesquisa, observando também a quantidade de citações. Além disso, considerando o arcabouço específico das leis brasileiras de licitações, foram levantados estudos realizados pelos órgãos de controle do país, de forma a coletar e analisar como o tema de detecção de indícios de irregularidades em compras públicas tem sido abordado pela CGU, TCU, Tribunais de Contas dos Estados e Receita Federal do Brasil. Por fim, foi realizada análise de co-citações de forma a buscar aqueles autores que formam a base do conhecimento sobre mineração de dados.

Para a construção do modelo preditivo, as tarefas de mineração seguem as etapas do modelo de referência CRISP-DM [11]: Entendimento do Negócio; Entendimento dos Dados; Preparação dos Dados; Modelagem; Avaliação; e Implantação, da seguinte forma:

- Na etapa de Entendimento do Negócio, aprofundou-se nas questões relevantes para a identificação dos indícios de irregularidades em licitações, bem como seus atores, arcabouço normativo e partes interessadas no Tribunal de Contas.

- Para a etapa de Entendimento dos Dados, avaliou-se as bases de dados existentes. Foram levantados os principais atributos que podem contribuir para classificar um certame como irregular.

- Na etapa de Preparação dos Dados, os atributos definidos foram carregados em base de dados relacional *MySQL* através de consulta em *SQL* e execução de processos de Extração, Transformação e Carga (ETL) com Python e auxílio de planilhas eletrônicas. O período de levantamento de licitações correspondeu a janeiro de 2014 a dezembro de 2019. Todas a etapas de processamento e análise foram feitas com *Python*, utilizando principalmente a biblioteca *scikit-learn*[9]. Métodos de seleção de atributos são aplicados, como *Recursive Feature Elimination* (RFE) e *Least Absolute Shrinkage and Selection Operator* com *Cross-Validation* (LASSOCV). Os dados foram separados em dados de treino/teste e uma amostra de avaliação.

- Na etapa de Modelagem selecionou-se o algoritmo de aprendizado de máquina mais adequado, com base no levantamento do estado da arte, através de treinamento supervisionado, com ajuste de parâmetros para otimização de cada modelo. Dado que o problema proposto trata-se da obtenção indicador contínuo para ranqueamento

[9]Disponível em: https://scikit-learn.org

com base em risco, a princípio cogitou-se uma abordagem de regressão, no entanto, foram utilizados algoritmos de classificação para obtenção das razões das chances da classificação binária para se estabelecer o *ranking*.

- Na etapa de Avaliação aplicou-se medidas de desempenho nos melhores modelos treinados para verificar sua confiabilidade.
- Por fim, na etapa de Implantação, o resultado da pesquisa foi posto para especialistas e uma estratégia de implantação foi apresentada.

1.4 Estrutura da Dissertação

Esta dissertação está organizada em cinco capítulos, incluindo este inicial no qual é feita uma introdução ao tema objeto de estudo, contendo a justificativa do tema escolhido e apresentação dos objetivos.

No Capítulo 2 apresenta-se o referencial teórico relacionado a licitações e controle externo. Em seguida fundamenta-se a mineração de dados, modelo de referência CRISP-DM, técnicas de tratamento e validação de dados e de avaliação de modelos, apresentando-se, ainda, trabalhos relacionados.

No Capítulo 3 é detalhada a metodologia aplicada neste estudo, que será orientada pelo modelo de referência CRISP-DM. No Capítulo 4 são apresentados os resultados da aplicação das técnicas discutidas. Além disso, são realizadas análises nos resultados encontrados, para verificar a qualidade e a viabilidade do modelo proposto.

Por fim, são apresentadas as conclusões e vislumbres de trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo são apresentadas as principais referências teóricas e trabalhos acadêmicos nos quais a solução proposta está embasada.

2.1 Licitações e o Controle Externo

2.1.1 Controle Externo e Tribunal de Contas do Estado de Goiás

Dentre as diversas facetas que o controle pode assumir no âmbito da administração pública, pode-se conceituá-lo de forma resumida como sendo uma ação fiscalizatória com o objetivo de examinar a gestão e aplicação dos recursos públicos em benefício da coletividade, incluindo a verificação da legalidade, legitimidade, economicidade e regularidade dos atos administrativos, que devem estar em conformidade com as normas constitucionais e infra-constitucionais [13, 9].

Conforme o posicionamento do agente controlador, o controle se divide em controle interno, controle externo e controle social [9]. No controle interno o agente controlador está inserido na estrutura do agente controlado. O art. 74 da Carta Magna obriga a todos os poderes a manutenção do sistema de controle interno integrado.

O controle externo é aquele exercido por um ente independente, localizado fora da estrutura de decisão do agente controlado, podendo ser executado pelo Poder Judiciário, pelo Poder Legislativo e pelos Tribunais de Contas [13]. O controle externo, pautado na necessidade da imposição de limites ao exercício do poder em favor do interesse público, fundamenta a existência dos Tribunais de Contas [13]. O controle social pode ser entendido como uma modalidade de controle externo em que o agente controlador é o cidadão [9].

O Tribunal de Contas do Estado de Goiás é órgão de controle externo que auxilia a Assembleia Legislativa na tomada de decisões técnicas. O TCE-GO é responsável por julgar contas de gestores e apreciar contas do governador; realizar auditorias de aspecto

contábil, orçamentário, financeiro ou patrimonial; realizar o controle de legalidade de atos de admissão de pessoal e de processos licitatórios de todos os Poderes [7]. Dentre as atribuições conferidas pelas normas, está a análise prévia de processos licitatórios.

A Figura 2.1 apresenta de forma simplificada o fluxo de um processo licitatório e os pontos de atuação do controle externo. Diante de uma necessidade de compra de bens ou serviços, pautada no interesse público, o governo produz um documento definindo as regras, prazos, preços, custos da contratação, que é chamado de edital de licitação. Nesse momento o Tribunal de Contas tem a prerrogativa de fiscalizar o processo de compra e todos os seus documentos, verificando o atendimento à legislação aplicável.

A partir de então, em regra, os particulares que têm interesse em contratar com o governo entram em disputa pela proposta que melhor atenda aos requisitos previstos no edital. Novamente, o Tribunal pode atuar para verificar a condução do processo licitatório.

Por fim, os particulares detedores das melhores propostas são declarados vencedores da licitação e, em regra, firmam um contrato ou instrumento semelhante com o governo para fornecer um produto ou serviço. O Tribunal de Contas tem prerrogativa de, a qualquer tempo, via denúncia, representação ou um instrumento de fiscalização, auditar o andamento do contrato.

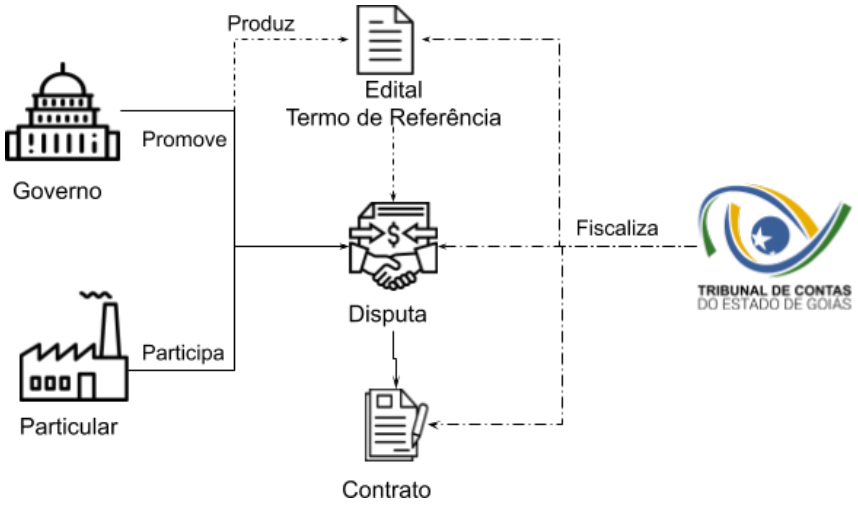

Figura 2.1: Atuação do Tribunal de Contas.

Fonte: Próprio autor

Para tanto, o TCE-GO possui duas áreas especializadas em análise prévia de licitações. A primeira é o o Serviço de Análise Prévia de Editais e Licitações, responsável pelo controle das compras de bens e serviços, tais como, medicamentos, tecnologia da informação, alimentos e serviços gerais contratados via terceirização. A segunda área é o Serviço de Análise de Editais e Projetos de Engenharia, área especializada em contratações de obras

e serviços de engenharia, tais como reformas, construção de obras rodoviárias e de obras civis.

Conforme o organograma do Tribunal [1], as áreas de análise de editais são parte da Secretaria de Controle Externo. Esta unidade, que é ligada diretamente à Presidência, é responsável pela execução das atividades finalísticas de controle externo.

Outro braço da Secretaria é o Serviço de Informações Estratégicas, unidade responsável por atividades de análise e cruzamento de dados massivos e participação em ações de fiscalização em conjunto com outros Tribunais de Contas ou órgãos de controle interno, como a CGU. Essa unidade possui acesso a diversas bases de dados de interesse do controle externo, tais como RAIS, Receita Federal, Tribunal Superior Eleitoral, Junta Comercial do Estado de Goiás (JUCEG).

Outra importante atribuição do Serviço de Informações Estratégicas é a implementação de rotinas automatizadas de auditoria, também conhecidas como trilhas de auditoria. As trilhas são construídas com bases em tipologias de irregularidades levantadas junto aos especialistas em fiscalização e tem o objetivo de aplicar técnicas computacionais e cruzamento de dados para detecção e envio de alertas automáticos para tomada de decisão.

2.1.2 Licitações

A administração pública tem o dever de observar os princípios da legalidade, impessoalidade, publicidade, moralidade e eficiência. Dessa forma, o art. 37 XXI da Constituição Federal determina, salvo casos ressalvados em lei, que os contratos públicos para obras, prestação de serviços, compras e alienações devem ser precedidos de processo licitatório, com vistas a assegurar a todos os interessados iguais condições de apresentar propostas e de serem escolhidos, desde que preencham os requisitos legais previstos. Além disso, segundo o art. 3 da Lei Federal 8.666/93, a licitação tem como objetivo selecionar a proposta mais vantajosa para a administração pública [14, 15].

A Lei Geral das Licitações (Lei Federal 8.666/93), em seu art. 22, e a Lei do Pregão Eletrônico (Lei Federal 10.520/02) definem as modalidades de licitação, que são podem ser consideradas formas específicas para condução do processo licitatório, cada uma com suas regras e procedimentos [16]. A concorrência é modalidade genérica, mais rigorosa em termos formalismo, aplicada a partir de critérios de valor e de natureza do objeto, em que participam quaisquer interessados que atendam aos critérios estabelecidos, sendo utilizada para contratações complexas e de grandes valores [14, 15, 16]. Em regra, conforme a Lei 8.666/93, a concorrência é obrigatória para contratos de obras e serviços de engenharia a partir de R$ 3,3 milhões ou para compra de bens ou serviços acima de R$ 1,4 milhões.

[1] Organograma do TCE-GO: https://portal.tce.go.gov.br/estrutura-organizacional

A modalidade tomada de preços é utilizada para valores intermediários entre a concorrência e o convite, sendo mais simples que a concorrência e com participação mais restrita, pois em regra, participam apenas licitantes previamente inscritos até o terceiro dia útil antes da licitação [15].

A modalidade convite é um procedimento simplificado de contratação, indicado para contratações de valores menores, em que a administração seleciona pelo menos três licitantes devidamente qualificadas (Súmula TCU 248) e envia para cada uma o instrumento convocatório chamado carta convite. O art. 22 da Lei Federal 8.666/93 define a modalidade leilão para alienação de bens imóveis, alienação de móveis inservíveis ou de produtos legalmente apreendidos ou penhorados, para aquele que ofertar o maior preço a partir de um preço de referência. Por fim, a lei das licitações define a modalidade concurso para escolha de trabalhos técnicos, artísticos ou científicos. Essas últimas não serão escopo deste trabalho.

A Lei Federal 10.520/02 definiu a modalidade pregão para aquisição de bens e serviços comuns, pelo menor preço, com padrões mínimos de qualidade mensuráveis e conhecidos no mercado e previstos no edital e termo de referência. No pregão não há limite de valor máximo para a sua utilização.

O pregão pode ser presencial, quando os licitantes devem comparecer a um lugar e hora previamente estabelecidos ou pode ser realizado na forma eletrônica, em que os licitantes acessam alguma plataforma na internet para participar da licitação. O Tribunal de Contas do Estado de Goiás tem decidido no sentido de que o gestor deve optar pelo pregão presencial apenas em casos em que não for possível a aplicação do pregão eletrônico (Acórdão 352/2015-TCEGO-Plenário). No Estado de Goiás, o Poder Executivo utiliza o Sistema Compras-NET para realização das licitações na modalidade pregão.

Por fim, na Lei 8.666/93 existem as hipóteses de compras diretas pelo poder público, exceções ao art. 37 XXI da Constituição Federal, através da inexigibilidade de licitação ou da dispensa de licitação. Conforme art. 25 da Lei 8.666/93, a inexigibilidade é aplicável apenas quando não há possibilidade de competição, ou seja, quando há fornecedor exclusivo ou quando o objeto da aquisição tem natureza singular ou a contratação é feita com profissionais ou empresas de notória especialização ou para contratação de profissional artístico consagrado. A dispensa de licitação é uma contratação direta em que existe a possibilidade de competição, mas ela é inoportuna diante do interesse público.

Além da Lei Geral da Licitação e da Lei do Pregão, o Estado de Goiás estabeleceu a Lei 17.928/12, que "dispõe sobre normas suplementares de licitações e contratos pertinentes a obras, compras e serviços, bem como convênios, outros ajustes e demais atos administrativos negociais no âmbito do Estado de Goiás".

A modalidade adotada na execução da compra pública pode ter relação com a ocorrência de irregularidades, pois algumas modalidades possuem ritos menos formais, menor transparência e concorrência reduzida ou ainda maior grau de discricionalidade do gestor governamental na tomada de decisão [17]. Nesse sentido, Rodidigues & Notato [16] verificaram a relação entre a modalidade de licitação e o risco de ocorrência de fraudes nos processos licitatórios realizados pelos municípios baianos, a partir de relatórios de controle interno da CGU entre 2004 e 2014. Os autores realizaram testes de hipóteses e demostraram que carta convite e dispensa de licitação apresentaram maiores percentuais de irregularidades, enquanto que concorrência e pregão apresentaram os menores indicativos.

Por fim, o arcabouço normativo aplicado às compras públicas é vasto e não é o objetivo deste trabalho esgotar este tema. Além das leis já apresentadas nesta seção, existem ainda diversas normas aplicáveis advindas de leis federais, estaduais e da jurisprudência. Assim, durante o desenvolvimento deste trabalho algumas normas adicionais serão apresentadas para dar suporte legal aos resultados apresentados, sem detalhamento do aspecto da doutrina de Direito.

2.1.3 Irregularidades em Licitações

As aquisições de bens e serviços públicos devem se submeter ao processo licitatório regido por leis e princípios constitucionais. Durante a execução das fases do certame, as irregularidades podem ocorrer desde o início, como por exemplo, a coleta incorreta de orçamentos ou adição de cláusulas que restringem a competitividade; até às fases finais, na adjudicação do objeto a um vencedor que não cumpre requisitos de idoneidade [18]. Os envolvidos em uma compra pública irregular podem ter a intenção de fraudar o processo licitatório, simulando a competição e direcionando a contratação para maximizar os seus lucros [19].

Identificar irregularidades nas licitações requer, por parte do auditor, a análise detalhada da legalidade da contratação, da aderência dos objetivos da licitação com os objetivos e necessidades da administração pública, verificação da compatibilidade de preços com valores de mercado, análise de documentos do processo licitatório e daqueles enviados pelos participantes. Dessa forma, cabe ao auditor a função de demostrar elementos comprobatórios para sustentar os apontamentos feitos, que derivam da comparação de um cenário esperado, com base em leis, normas e/ou boas práticas, com o cenário efetivo encontrado [19].

As evidências de auditoria são elementos constatados pelo auditor que fundamentam sua opinião de certeza, comprovam uma situação de fato e que podem ser obtidas por processos analíticos em demostrações financeiras e lançamentos contábeis, através análise

de documentos comprobatórios ou por inspeção física. O indício é uma uma constatação de uma discrepância entre a situação encontrada e a situação esperada, mas sem elementos diretos capazes de emitir uma opinião com grau de certeza [20]. No entanto, segundo a jurisprudência do Supremo Tribunal Federal (STF), a existência de um conjunto de indícios é uma evidência (RE nr. 68.006-MG), dessa forma, um conjunto de indícios de irregularidades que tenham correlação com fato em análise pode ser utilizado em processos de auditoria por meio de evidência indireta (Acórdão 630/2006-TCU-Plenário).

A Tabela 2.1 apresenta alguns exemplos de tipologias de indícios de irregularidades em licitações encontradas na literatura e a base legal a qual se relaciona [16, 18, 21, 19].

Tabela 2.1: Irregularidades em licitações

Irregularidade	**Base Legal**
Ausência de comprovação de capacidade técnica	Art. 30 da Lei 8.666/93
Ausência de declaração do ordenador de despesa	Art. 16 da Lei 101/01;
Ausência de orçamento detalhados e preços unitários	Art. 7 e art. 40 da Lei 8.666/93
Ausência de programação de desembolso financeiro	Decretos estaduais que estabelecem normas complementares de programação e execução orçamentária, financeira e contábil para o exercício
Ausência de projeto básico e orçamentos em planilhas	Art. 7 e Art. 40 da Lei 8.666/93; Art. 12 da Lei Estadual 17.928/12;
Ausência ou aplicação incorreta de BDI	Súmula 253 TCU; Acórdão 2622/2013-TCU-Plenário
Ausência ou deficiência de pesquisa de preços	Art. 15 e art. 43, IV, da Lei 8.666/93; Acórdão 158/2019-TCEGO-Plenário; Acórdão 183/2019-TCEGO-Plenário
Ausência ou deficiência na publicidade e transparência	Art. 3 e art. 21 da Lei 8.666/93

Tabela 2.1 continuação da página anterior

Irregularidade	**Base Legal**
Ausência ou deficiência no planejamento prévio da contratação	Art. 1, Lei 101/01; Art. 6 do Decreto-Lei 200/07; Art. 174 Constituição Federal; Art. 3, I, da Lei do 10.520/02; Art. 11 Lei Estadual 17.928/12; Acórdão 2640/2018-TCEGO-Plenário; Acórdão 2814/2018-TCEGO-Plenário
Autor do projeto e o licitante vinculados	Art. 9 da Lei 8.666/93
Composição da comissão de licitação irregular	Art. 51, da Lei 8.666/93; Art. 1 e 3 da Lei 10.520/02;
Conluio	Art. 3 da Lei 8.666/93
Desclassificação/Inabilitação indevida de licitante	Art. 48, II, alíneas A e B da Lei 8.666/93; Acórdão 3.381/2013-TCU-Plenário; Acórdão 11.907/2011-Segunda Câmara TCU; Acórdão 2.162/2018-TCEGO-Plenário
Despesas realizadas sem licitação	Art. 37 XXI da Constituição Federal
Direcionamento de contratação	Art. 3 da Lei 8.666/93
Dispensa ou inexigibilidade sem fundamentação legal	Art. 37 XXI da Constituição Federal
Fracionamento de despesa	Art. 23 da Lei 8.666/93
Inobservância da previsão de preferência a contratações de ME e EPP	Lei Federal 123/2006
Não parcelamento de objeto	Art. 23 Lei 8.666/93; Súmula 247 TCU
Objeto impreciso, genérico, incompreensível ou incompleto	Art. 40 e art. 47 da Lei 8.666/93
Outras irregularidades	Irregularidades que violam diversos princípios e normas do direito
Previsão de Sub-contatação irregular	-
Restrição de competitividade - Cláusulas restritivas	Art. 3 e art. 30 da Lei 8.666/93; Acórdão 461/2014-TCU-Plenário
Superfaturamento ou sobrepreço em orçamentos	Art. 96 da Lei 8.666/93
Uso de modalidade indevida	Art. 5 da Lei 12.232/2010

Tabela 2.1 continuação da página anterior

Irregularidade	**Base Legal**
Uso indevido de Pregão Presencial em detrimento de Pregão Eletrônico	Art. 85 da Lei Estadual 17.928/2012; Acórdão 352/2015-TCEGO-Plenário
Utilização de critérios de habilitação ou julgamento sem previsão em edital	Art. 3 e art. 40 da Lei 8.666/93
Vínculos entre licitantes e servidores públicos	Art. 9 Lei 8.666/93; Acórdão 1198/2007-TCU-Plenário

2.2 Mineração de Dados

O trabalho proposto utiliza mineração de dados em benefício do controle prévio e concomitante nas compras governamentais. A mineração de dados tem como objetivo a busca de padrões que possam prever comportamentos a partir de análise dados históricos [22]. Para tanto, utiliza aprendizado de máquina através do treinamento de um algoritmo capaz de processar tais dados de três formas: supervisionada, não supervisionada e por reforço [23].

Na aprendizagem supervisionada, os dados de treino já possuem uma classe previamente conhecida, uma medida de adequação para classificação dos dados com a qual se pode validar o resultado. Por outro lado, na aprendizagem não supervisionada o próprio algoritmo deve buscar explicar os resultados [24].

Na aprendizagem por reforço, o algoritmo aprende explorando as possibilidades do ambiente ao qual se insere, buscando, por tentativa e erro, quais são as possibilidades que tem maior recompensa ou menor penalidade [24].

As tarefas de aprendizado de máquina são divididas em quatro grupos: classificação, regressão, agrupamento e regras de associação [22]. As tarefas de classificação caracterizam um atributo da relação com um resultado nominal ou categórico. As tarefas de regressão fazem o mesmo, mas o resultado é um valor numérico. As tarefas de agrupamento criam grupos de instâncias com características comuns. E as tarefas de associação identificam regras que associam conjuntos de itens.

2.2.1 Tarefas de Regressão

Modelos de regressão são comumente utilizados quando a variável de resposta é do tipo quantitativa ou em problemas binários [25]. Os modelos baseados em regressão buscam prever uma variável aleatória dependente Y a partir de uma ou mais variáveis aleatórias independentes $X^t = (X_1, X_2, X_3..., X_p)$ [26].

O modelo mais simples é a regressão linear, em que se pressupõe uma relação linear entre Y e X^t, com coeficientes β_j desconhecidos, segundo uma função de predição $f(X)$ dada pela Equação 2.1 [26]:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{2.1}$$

Por outro lado, quando a relação entre Y e X_t não é linear, outras técnicas devem ser aplicadas para capturar a não linearidade. Nesta esteira, os Modelos Lineares Generalizados (MLG) são uma extensão dos modelos de regressão simples e múltipla, permitindo modelar variáveis de interesse na forma de contagem, contínuas, binárias ou categóricas.

A Regressão Logística faz parte dos Modelos Lineares Generalizados para variável de interesse binária contida no intervalo de zero a um, devido ao uso da função *logit* da Equação 2.2.

$$p(x) = \frac{e^(\beta_0 + \beta_1 x_1 + \cdots + \beta_s x_s)}{1 + e^(\beta_0 + \beta_1 x_1 + \cdots + \beta_s x_s)} \tag{2.2}$$

A Equação 2.2 assume valores entre $[0, 1]$ para todo x. Para este trabalho a variável dependente com valor mais próximo a 1 indica a presença de irregularidade em uma licitação. O valor 0 indica ausência de irregularidade.

2.2.2 Tarefas de Classificação

As tarefas de classificação pertencem ao grupo de aprendizado supervisionado em que, a partir de um conjunto de dados rotulados previamente, um modelo matemático é treinado para classificar novas informações ainda não rotuladas. São exemplos de algoritmos de classificação: *Gradient Boosting Machine* (GBM), *Random Forest* (RF) e *Support Vector Machine* (SVM).

***Gradient Boosting Machine* (GBM)**

Gradient Boosting Machine (GBM) é um algoritmo iterativo que combina um estimador fraco (alto erro de predição) para produzir um estimador altamente preciso, através da redução de uma função de perda [27, 28]. Neste caso, o *boosting* consiste em adicionar novos modelos ao conjunto sequencialmente, mas de forma diferente dos métodos tradicionais de *boosting* que ponderam amostras positivas e negativas, pois o *Gradient Boosting* para classificação (GBM) faz a convergência global do algoritmo seguindo a direção do gradiente negativo [29].

O GBM requer pouco pré-processamento de dados e ajuste dos parâmetros, é robusto para um pequeno conjunto de dados e pode ser aplicado a problemas de classificação ou

regressão. As interações complexas são modeladas de forma simples, os valores ausentes nos preditores são gerenciados quase sem perda de informações [28]. Existem diversas implementações deste algoritmo e neste trabalho é utilizada a versão "XGBM" desenvolvida por Tianqi Chen[2], que utiliza um *ensemble* de árvores de decisão como estimadores fracos.

Random Forest

As árvores de decisão são estruturas simples, flexíveis e de fácil interpretação, mas uma árvore utilizada isoladamente pode ser instável e susceptível a superajuste aos dados (*overfitting*) [30].

Breiman [31] propôs *Random Forest*, um algoritmo que combina (*ensemble*) uma coleção de árvores de decisão para uma tarefa de classificação ou regressão. O conjunto de dados é divido aleatoriamente em árvores de decisão independentes, formadas sem poda, através de *bootstrap aggregating* (*bagging*). Cada árvore vota em uma classe e a classe mais votada corresponde à resposta final [31, 25]. Por essas características, *Random Forest* é considerado um algoritmo estável, resistente ao superajuste, mais tolerante a ruídos e que lida bem com classes desbalanceadas [30].

Support Vector Machine (SVM)

Support Vector Machine (SVM) é um modelo de classificação largamente utilizado em problemas com dados lineares ou não lineares, tais como processamento de texto, de imagens e reconhecimento de padrões [32, 33].

SVM mapeia p atributos como vetores em um espaço p-dimensional e tenta encontrar um hiperplano ótimo que separa as classes, com uma margem de separação definida por vetores de suporte. Por exemplo, a Figura 2.2 apresenta uma classificação com SVM em um espaço bi-dimensional com um conjunto de dados com duas classes.

As observações que ficam diretamente na margem ou dentro dela são conhecidas como vetores de suporte [25]. Segundo [32] a construção hiperplanos ótimos, que leva em consideração uma pequena quantidade vetores de suporte dos dados de treinamento, permite alta capacidade de generalização, mesmo em um espaço dimensional infinito.

Quando os dados de treinamento não são linearmente separáveis, o SVM linear perde poder de generalização. Para resolver essa limitação, foi criado o "truque do kernel", em que o espaço dimensional original é mapeado em uma espaço de alta dimensão [25]. Existem vários tipos de *kernel* $K(x_i, x_k)$, tais como o *kernel* linear da Equação 2.3 para

[2]Fonte: https://xgboost.readthedocs.io/en/latest/tutorials/model.html, acessado em 10/11/2020

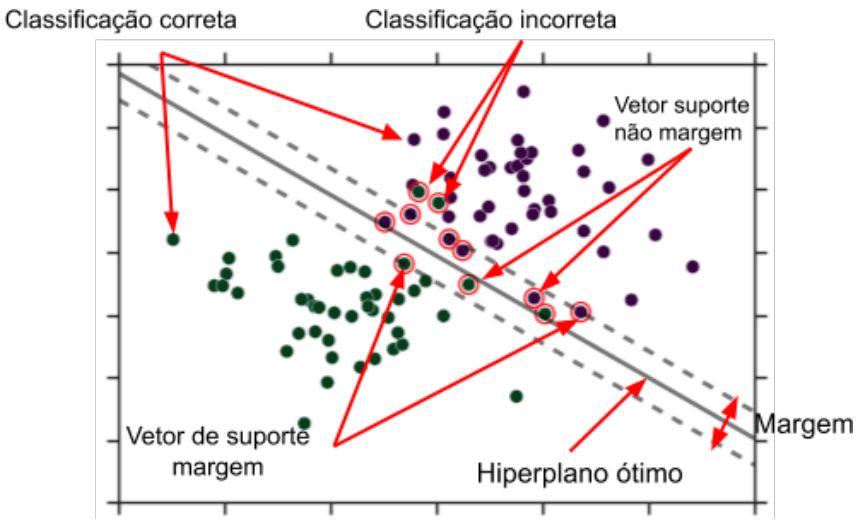

Figura 2.2: Classificação binária SVM com duas variáveis apresentado hiperplano ótimo, margem e vetores de suporte.

Fonte: Próprio autor

problemas lineares, o *kernel* polinomial da Equação 2.4 com grau d sendo um número inteiro e o *kernel* radial da Equação 2.5 com γ sendo uma constante [25].

$$K(x_i, x_k) = \sum_{j=1}^{p} x_{ij} x_{kj} \tag{2.3}$$

$$K(x_i, x_k) = (1 + \sum_{j=1}^{p} x_{ij} x_{kj})^d \tag{2.4}$$

$$K(x_i, x_k) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} x_{kj})^2) \tag{2.5}$$

A otimização do SVM pode ser feita com *cross-validation* ajustando-se o parâmetro C, que indica a tolerância de violações de vetores na margem ou ao próprio hiperplano. Por exemplo, na Figura 2.2, quanto maior o valor de C, maior a margem e, portanto, maior tolerância à violações [25].

2.3 Modelo de Referência CRISP-DM

A criação do modelo se baseia nas fases do CRISP-DM [11]. As etapas são as seguintes: Entendimento do Negócio; Entendimento dos Dados; Preparação dos Dados; Modelagem; Avaliação; e Implantação, conforme apresentado na Figura 2.3.

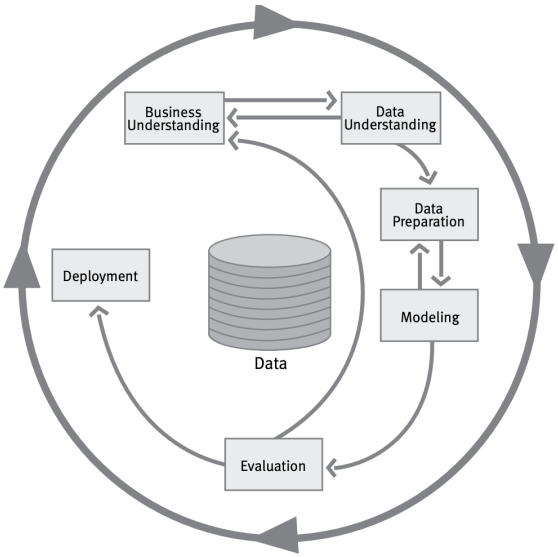

Figura 2.3: Fases do modelo CRISP-DM.

Fonte: [11, p. 10]

Entendimento do negócio

O entendimento do negócio é a fase inicial do projeto, similar ao processo de engenharia de requisitos na engenharia de software, ou seja, o foco é entender as necessidades do negócio apresentado sob a ótica do cliente e definir um plano de ação. Essa etapa tem quatro tarefas: determinar os objetivos do negócio; avaliar a situação; definir os objetivos da mineração e produzir um plano de ação.

Entendimento dos dados

Nesta fase estuda-se inicialmente os dados disponíveis, realizando amostragens para verificar tamanho, variedade, quantidade, qualidade, disponibilidade, dentre outras características relevantes. A visão inicial permite conhecer o universo de dados e suas potenciais fontes, permitindo detectar problemas, levantar hipóteses iniciais sobre a tarefa de mineração e encontrar padrões escondidos, por exemplo. Nessa etapa estão previstas as tarefas de coleta de inicial de dados; documentação inicial do dados; exploração; e verificação da qualidade.

Preparação dos dados

Nesta etapa o conjunto de dados para mineração é construído para utilização nas etapas posteriores. Estão inclusas tarefas de seleção de dados; limpeza de dados; construção de dados; integração; e formatação.

Esta etapa consome boa parte do trabalho de mineração, pois muitas vezes os dados não estão organizados e dessa forma são necessárias diversas técnicas limpeza, transformação, balanceamento de dados, organização e seleção de variáveis.

Em tarefas de aprendizado supervisionado, o conjunto de dados final é dividido em dados de treino/teste e dados de avaliação. O primeiro conjunto é utilizado para treinar o algoritmos de mineração na etapa de modelagem, enquanto que o segundo conjunto é utilizado para avaliar o modelo.

Modelagem

A partir do conjunto de dados organizado, são aplicados as tarefas de mineração. Em muitos casos mais de uma tarefa pode ser aplicada, variando-se os parâmetros de ajustes para buscar aquela com melhor resultado. Em outros, casos pode ser necessário retornar as etapas anteriores para ajustes nos dados.

Avaliação

Após a construção de um ou mais modelos, um conjunto de análises e comparações indicam qual é o melhor modelo que atende as necessidades do negócio. Pode ser necessário voltar as etapas anteriores para ajustes. Nessa etapa, decide-se se o produto da mineração deve ser ajustado, descartado ou ir para a etapa de implantação.

Implantação

Nesta fase, o produto da mineração é apresentado para as áreas de interesse, gerando valor para o negócio. O resultado pode ser uma aplicação perene de mineração dentro de um processo de negócio ou pode ser um relatório pontual e específico que apoia a tomada de decisão.

2.4 Tratamento e Validação de Dados

2.4.1 Balanceamento de Dados

Quando se cria a massa de dados de treinamento do modelo, os dados podem estar desbalanceados, isto é, com mais registros de uma classe de dados do que de outra. Para resolver esse problema, os algoritmos de balanceamento *under-sampling* e *over-sampling* podem ser aplicados. O primeiro remove registros da classe majoritária, enquanto o segundo replica registros da classe minoritária [34, 35].

A técnica *Synthetic Sampling with Data Gereration* (SMOTE) gera sinteticamente novas instâncias da classe minoritária, através de interpolação de valores. A técnica consiste em, a partir do conjunto de instâncias da classe minoritária C_m, selecionar uma instância aleatória $x_i \in C_m$ e calcular o conjunto $M \in C_m$ como sendo os K-vizinhos mais próximos de x_i, utilizando a menor distância euclidiana entre x_i e x_k, com $x_k \in C_m$ e K um número inteiro. A nova instância $x_n \in C_m$ é calculada conforme a equação 2.6, com δ sendo um número entre 0 e 1 [36, 35].

$$x_n = x_i + (x_k - x_i)\delta \tag{2.6}$$

2.4.2 Reamostragem

Nos modelos de aprendizagem estatística, quando se tem um grande conjunto de dados, esses podem ser divididos em duas partes, sendo uma para treino e outra para teste [26, 25]. Essa técnica é chamada de *holdout*. Nesse caso os dados de treino são utilizados para aprendizagem do modelo e os dados de teste são utilizados para avaliar a qualidade do modelo. A lógica dessa abordagem está em utilizar dados que o modelo treinado ainda não conhece (dados de teste) e a partir daí inferir o quão assertivo ele é.

No entanto, quando exitem poucos dados, essa divisão em treino e teste pode não ser possível, sendo necessário lançar mão de técnicas estatísticas específicas de reamostragem. Dentre essas, as mais utilizadas são *boostrap* e *cross-validation* [25].

O uso de reamostragem permite realizar inferências de um parâmetro populacional a partir de amostras geradas repetidas vezes nos dados originais [37]. Isto pode ser computacionalmente custoso, pois para cada amostra gerada um modelo de aprendizado deve ser treinado [25].

Bootstrap

Bootstrap é uma técnica computacional intensiva de reamostragem que permite obter estimativas de parâmetros e de erros desses parâmetros de uma população através de sucessivas amostragens em uma amostra inicial [38]. Esta técnica consiste em obter conjuntos de dados de treino a partir de reamostragem aleatória de mesmo tamanho da amostra original com repetição [26]. Quando a distribuição dos dados é conhecida, trata-se do *bootstrap* paramétrico e quando não se conhece a distribuição dos dados, trata-se do *bootstrap* não paramétrico [37].

No *bootstrap*, uma amostra inicial é feita na população. A partir dessa amostra, diversas outras são produzidas e uma distribuição empírica de uma variável de interesse θ é gerada sem necessariamente se conhecer qualquer distribuição prévia dessa variável na população. Assume-se que a distribuição empírica é próxima da distribuição da população

e, dessa forma, é possível estimar parâmetros complexos da população a partir dessa distribuição [26, 39, 38].

Cross-validation

A técnica de *cross-validation* consiste em utilizar parte dos dados para treino e parte para teste, agrupando esses dados em conjuntos de tamanhos aproximadamente iguais a partir de amostragem aleatória. A quantidade de conjuntos de dados é dada pela variável k ou $k - fold$ [26].

Quando k tem o mesmo tamanho n da amostra, a técnica recebe nome de *leave-one-out cross validation* [26, 25]. Dessa forma, uma observação aleatória (x_1, y_2) é utilizada para teste enquanto todas as demais observações $(x_2, y_2), (x_3, y_3)...(x_n, y_n)$ são utilizadas para treino. No entanto, essa técnica pode ser computacionalmente demorada, pois quanto maior o tamanho n do conjunto de dados, mais modelos são treinados e testados [26].

2.4.3 Padronização e Variáveis *Dummy*

É comum que em bases de dados reais alguns atributos tenham variações de magnitude. Nesses casos, pode ser necessário aplicar técnicas de padronização nos dados, fazendo com que os atributos permaneçam na mesma escala [40, 41]. O pacote *sklearn.preprocessing* fornece métodos para essa tarefa como o *StandardScaler*, que aplica padronização do atributo com relação a média μ e o desvio padrão σ, conforme equação 2.7.

$$z = \frac{x - \mu}{\sigma} \tag{2.7}$$

A maioria dos modelos de aprendizado de máquina lidam com dados exclusivamente numéricos. Nesses casos, os atributos categóricos devem ser transformados de forma que cada valor distinto do atributo k se torne um novo atributo, aumentando a quantidade de atributos final em $k - 1$, contendo apenas dois valores possíveis: zero (0), indicando ausência de valor na variável original ou um (1), indicando a presença de valor [42].

2.4.4 Seleção de Variáveis

Quando existem muitas variáveis independentes, algumas delas podem ser pouco importantes para o processo de classificação ou regressão, adicionando ruídos aos modelos, aumentando a complexidade computacional ou causando *overfitting* nos modelos [40]. Além disso, modelos mais simples tendem a ser mais interpretáveis e, no caso especifico deste trabalho, que utiliza variáveis de diversas bases de dados, quanto menos variáveis forem

necessárias para obtenção do resultado, mais simples e barato se torna a implantação da solução proposta no ambiente de produção.

Alguns atributos pouco importantes podem ser identificados através da análise de variância e de correlação. Atributos com variância igual a zero não contribuem com informações úteis, bem como atributos com baixo coeficiente de variação com relação a cada classe. O coeficiente de variação cv é uma medida do desvio padrão S pela média amostral $\bar{x}$, conforme Equação 2.8 [43].

$$cv = \frac{S}{\bar{x}} \tag{2.8}$$

Os atributos perfeitamente correlacionados podem ser considerados redundantes e também podem ser eliminados [40, 44]. O processo de eliminação deve ser supervisionado por um especialista de negócio.

Ridge, *Least Absolute Shrinkage and Selection Operator* (LASSO) e Elastic-Net

Nos métodos de regressão, o vetor de coeficientes é definido por alguma função, como por exemplo, método dos mínimos quadrados na regressão linear, em que se mantém todas as variáveis independentes no modelo final [26]. Contudo, quando se tem muitas variáveis, pode ser necessário reduzir aquelas menos significativas através de duas técnicas de regularização comumente utilizadas: *Ridge* e LASSO [44].

Ambas as funções utilizam uma penalização do vetor de coeficientes, de forma que quanto maior os coeficientes estimados, maior a penalização. *LASSO* utiliza a penalização $L1$ conforme 2.9, que é calculada como o somatório do módulo dos coeficientes. *Ridge* utiliza penalização $L2$ calculada através da soma dos quadrado dos coeficientes [26]. A aplicação de LASSO é indicada para redução de variáveis, pois tem a capacidade de zerar alguns coeficientes, removendo-os da resposta final. Para coeficientes correlacionados, LASSO tende a selecionar apenas um deles e zerar os demais. *Ridge* reduz os coeficientes das variáveis, mas não os zera.

Na Equação 2.9 e Equação 2.10, a variável β_j representa o *j-esimo* coeficiente, p indica o número de coeficientes e a variável λ é um ajuste de impacto da regularização.

$$L1 = \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.9}$$

$$L2 = \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.10}$$

A penalização *ElasticNet* utiliza uma combinação de *Ridge* e LASSO para selecionar variáveis [26, 45]. A Equação 2.11 define a penalização *ElasticNet*. Ela utiliza os parâmetros α e λ para controle da penalização. O parâmetro α assume valores entre $[0, 1]$ e controla a penalidade $L1$ e $L2$. Observando a Equação 2.11, percebe-se que se $\alpha = 0$, aplica-se apenas a penalização *Ridge* apresentada na Equação 2.9, mas se $\alpha = 1$, aplica-se apenas a penalização LASSO apresentada na Equação 2.10. O parâmetro λ controla o impacto da regularização.

$$E = \lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1 - \alpha)|\beta_j|) \tag{2.11}$$

Recursive Feature Elimination RFE

O método RFE é um processo que utiliza algum modelo de classificação para recursivamente treiná-lo e elaborar um ranking dos atributos, segundo algum critério obtido do algoritmo de classificação. A cada iteração, as variáveis menos importantes são removidas do ranking [46].

O RFE pode ser utilizando com uma Regressão Logística (RL) com penalização $L2$ em um conjunto de dados de interesse. Nesse caso, a cada iteração o algoritmo de RL é treinado e os atributos ordenados pelos coeficientes obtidos na regressão. Os atributos com menores coeficientes são eliminados. Para melhorar a sua assertividade, pode-se realizar o processo de treino com *cross-validation*.

Least Angle Regression - Lars

Least Angle Regression (Lars) é um método de seleção de variáveis que inicia o processamento com todos os coeficientes zerados e calcula o valor residual r ($r = y - \bar{y}, \beta_1, \beta_2..., \beta_p = 0$) sendo p o número de variáveis. Em seguida, seleciona a variável x_j com maior correlação com r e move seu coeficiente β_j continuamente em direção ao valor de mínimos quadrados até que outra variável x_k atinja correlação com valor residual r. A variável x_k é capturada e o processo se repete até que o número de preditores termine [47, 26, 48].

Seu funcionamento é semelhante ao *Forward Stepwise*, mas tem desempenho superior pela condição de parada (encontro da variável com maior correlação), pois executa m passos até completar a tarefa de seleção, sendo m o número de co-variáveis [47, 26].

Principal Component Analysis (PCA)

A análise de componentes principais é uma técnica aplicada para o entendimento de um conjunto de dados, para a redução de dimensionalidade, detecção de *outliers* ou seleção de

variáveis [49]. Trata-se de uma técnica não supervisionada que utilizada apenas as variáveis independentes $x_1, x_2...x_p$ para reduzir a dimensão p dos dados em n componentes que capturam a maior parte da variação existente nos dados originais, através de combinações lineares. O primeiro componente captura a maior variabilidade e assim sucessivamente [25, 50].

2.5 Técnicas de Avaliação de Modelos

As técnicas de avaliação utilizadas se baseiam na matriz de confusão da Tabela 2.2 [51, 52].

Tabela 2.2: Matriz de Confusão

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

Os valores da Tabela 2.2 são explicados no contexto de irregularidades em licitações da seguinte maneira:

- Verdadeiro Positivo (VP): licitação classificada pelo modelo como com alto risco de irregularidade e que realmente era de alto risco.
- Falso Positivo (FP): licitação classificada pelo modelo como com alto risco de irregularidade, mas que na verdade é de baixo risco.
- Falso Negativo (FN): licitação classificada pelo modelo como com baixo risco de irregularidade, mas que de fato é de alto risco.
- Verdadeiro Negativo (VN): licitação classificada pelo modelo como com baixo risco de irregularidade e que realmente é de baixo risco.

2.5.1 Medidas de Validação

Acurácia

Mede o percentual de acertos do modelo.

$$Acurácia = \frac{VP + VN}{VP + FP + FN + VN} \tag{2.12}$$

Sensibilidade

A sensibilidade mede a proporção entre o total de originalmente positivos observados e os valores corretamente previstos como positivos [53, 52, 54].

$$Sensibilidade = \frac{VP}{FP + FN} \tag{2.13}$$

Precisão

A precisão mede a acurácia do modelo em termos de acertos positivos [53, 52, 54].

$$Precisão = \frac{TP}{TP + FP} \tag{2.14}$$

Especificidade

A especificidade mede a proporção entre o total de originalmente negativos observados e os valores corretamente previstos como negativos [53, 52, 54].

$$Especificidade = \frac{VN}{FP + VN} \tag{2.15}$$

F-mesure

A *F-measure* é uma métrica que incorpora a Sensibilidade e a Precisão do modelo [53, 52, 54].

$$F\text{-}measure = \frac{(1 + \beta).Precisão.Sensibilidade}{\beta^2.(Precisão + Sensibilidade)} \tag{2.16}$$

O parâmetro β permite dar peso maior à Sensibilidade quando $\beta > 1$ ou maior Precisão quando $0 < \beta < 1$.

Área sob a Curva ROC (AUROC)

Receiver Operating Characteristic (ROC) é uma curva em um gráfico bidimensional que representa a taxa de verdadeiros positivos no eixo Y e a taxa de falsos positivos no eixo X [55, 36, 56]. Na Figura 2.4 a linha azul representa a Curva ROC. A linha diagonal em vermelho representa a estratégia aleatória de classificação. Classificações abaixo da diagonal indicam um classificador ruim.

A área sob a cuva ROC (AUROC) é um medida que varia entre 0 e 1 e permite comparar o desempenho de classificadores. Quanto mais próximo de 1, melhor é o classificador, pois indica que ele tem a taxa de verdadeiros positivos maior do que a taxa de falsos positivos [55, 54].

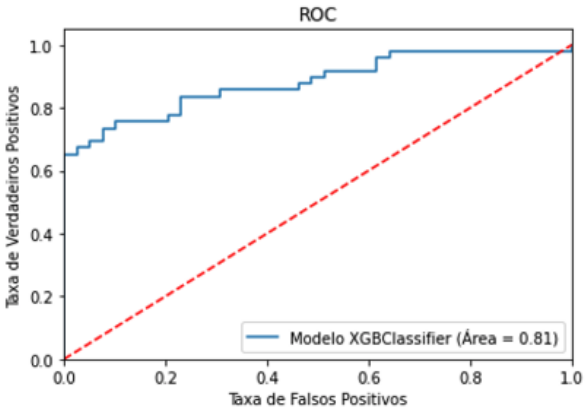

Figura 2.4: Curva ROC para modelo XBG

Fonte: Próprio autor

2.6 Trabalhos Relacionados

Nos últimos três anos foram encontradas pesquisas no campo de aprendizagem de máquina associadas à classificação e predição de corrupção [57, 58], lavagem de dinheiro [59], no combate a fraudes bancárias [60, 61], fraudes em cartões de crédito [41, 62, 63, 64] telecomunicações [42]; cripto-moedas [65], comércio eletrônico [66], seguros [67], na avaliação de risco de empresas e ativos [68].

Nos trabalhos pesquisados, observou-se a tendência de uso de técnicas de classificação com SVM [61, 60, 58, 69, 41, 70], *Random Forest* [61, 41, 70, 69], Regressão Logística [71, 69, 41], *Gradient Boosting* [72, 69, 73] e redes neurais [70, 74, 60, 57]. Outros trabalhos fizeram uso de *Naïve Bayes* [75, 62, 71, 64]. Em outros casos, árvores de decisão foram aplicadas para tarefas de classificação [58, 63, 76, 77], bem como técnicas de regras de associação e clusterização. Portanto, é vasta a quantidade de técnicas utilizadas pelos pesquisadores, com atenção para tarefas supervisionadas de classificação para determinar cenários de fraudes e irregularidades.

O trabalho Iturriaga [57] propõe o uso de uma rede neural *Self-Organizing Maps - SOMs* - para predizer o risco de corrupção com base fatores sociais, políticos e econômicos em províncias da Espanha. O autor utilizou uma base de dados de notícias na internet para coleta de casos de corrupção. Essas informações foram correlacionadas com outras variáveis socioeconômicas, políticas e geográficas para treinamento supervisionado de uma *SOM*.

Taha e Malebary [69] utilizaram técnicas de classificação para detecção de transações fraudulentas em cartões de crédito. A proposta dos autores foi estudar a aplicação de *Gradient Boosting* em uma base de dados com 284.807 transações de cartão de crédito

feitas na Europa, através da otimização de hiper parâmetros no classificador *LightGBM*. O resultado foi comparado com outros classificadores, como SVM, *Random Forest*, Regressão Logística dentre outros. Os autores concluíram que *LightGBM* otimizado apresentou desempenho superior aos classificadores tradicionais. Eles ainda reforçaram a importância do ajuste de parâmetros para incremento das medidas de desempenho dos algorítmos.

Com relação a investigação de irregularidades em licitação, tema do trabalho proposto, foram encontrados estudos produzidos por pesquisadores ligados à órgãos públicos de controle interno ou externo no Brasil, como a Controladoria Geral da União e o Tribunal de Contas da União.

O trabalho de Sarmento [77] utilizou técnicas de regras de associação e clusterização para identificar a formação de cartéis em licitações. Foi utilizado o algoritmo *Apriori* para obtenção de regras de associação entre empresas que participavam de licitações no governo federal. Foi utilizado o algoritmo *EM* (*Expectation-Maximization*), baseado em *K-means*, para clusterizar as empresas por área de geográfica de atuação.

O estudo Carvalho et al. [78] realizou comparativo de algoritmos de Redes Bayesianas para investigação de fraude de fracionamento de licitação nas compras do Governo Federal do Brasil. O estudo percorreu todas as etapas do modelo CRISP-DM [11] para produção de conhecimento, detalhando o trabalho realizado em cada uma. Na etapa de modelagem, por exemplo, foi utilizado o software livre *Weka* com ajustes de hiper parâmetros para cada algoritmo que implementa o modelo de aprendizagem. Foi feito um comparativo do desempenho e efetividade de diferentes algoritmos que implementam *Naïve Bayes* e *Baeysian Network*, usando *10-fold cross-validation* para melhoria da acurácia do modelo. Adicionalmente, para estudar a influência das classes desbalanceadas sobre o resultado, todos os algoritmos foram treinados com massas de dados desbalanceadas; com dados balanceados com *unsersampling* apenas; com *oversamping* apenas; e com *unsersampling* e *oversamping* ao mesmo tempo.

O trabalho de Carvalho & Carvalho [44] apresentou estudo de aplicação de Redes Bayesianas para criar um indicador de risco de fraude de corrupção em unidades administrativas do Governo Federal no Brasil. Os autores realizaram um processo de seleção de variáveis com regressão por regularização utilizando *Adaptative Lasso*. Realizaram discretização com *Minimum Description Length Principle* (MDLP) e com *Class-Attribute Contingency Class* (CACC), pois alguns dos modelos treinados necessitavam de dados categóricos. O estudo concluiu que a discretização com MDLP demostrou ser mais eficiente em termos de consumo de recursos do que a realizada com CACC. Por fim, os autores compararam o desempenho de três modelos (*Naive Bayes*, *Tree Augmented Naïve Bayes* e *Attribute Weighted Naïve Bayes*), sendo que *Naïve Bayes* demostrou melhores resultados do que os demais.

O trabalho de Balaniuk [79] aborda a fraude no serviço público inicialmente de forma genérica, identificada por padrões que fogem da normalidade dos atos administrativos. Num escopo mais específico, o estudo avalia riscos nos contratos entre o governo e os particulares. O artigo descreve o método de treinamento supervisionado para classificar eventos em alto ou baixo grau de risco de fraude. Nesse estudo, a resposta do modelo é baseada na probabilidade condicional de sete atributos capturados a partir da discretização de fatores de riscos apontados por especialistas do negócio. O autor chama a atenção para o problema da qualidade dos dados gerados pelo serviço público, alertando para o fato de que a ocorrência da fraude pode ser considerada uma classe rara e que, na maioria das vezes, permanece armazenada em documentos impressos.

Além desses, foram encontrados estudos que utilizaram aprendizado não supervisionado com redes neurais profundas (*Deep Neural Networks*) *Autoencoders* para detecção de anomalias. Paula *et al.* [59] realizaram estudo comparativo entre *Principal Component Análise* (PCA) e rede *Autoencoder* para detectar indícios de lavagem de dinheiro através da detecção de anomalias, processando cerca de 819.990 registros de exportações de empresas brasileiras coletados no ano de 2014. O estudo concluiu que a rede *Autoencoder* é cerca de 20 vezes mais rápida para redução de dimensionalidade, indicando ainda que essa rede é mais adequada para generalizações de problemas não lineares. Além disso, os resultados demostraram a viabilidade de se utilizar a rede neural em problemas reais para detecção de lavagem de dinheiro na Receita Federal do Brasil.

Em estudo similar, Domingos *et al.* [80] utilizaram treinamento não supervisionado com redes *Autoencoders* para detecção de fraudes em licitações, através da detecção de anomalias em cerca de 137.035 registros e 31 atributos de aquisições de tecnologia da informação do Governo Federal do Brasil. O estudo realizou comparação de desempenho da rede neural com diferentes parâmetros (número de épocas, número de camadas ocultas, diferentes funções de ativação), utilizando como medida o Erro Quadrático Médio (MSE). O modelo treinado com 5 épocas, 15 camadas ocultas e função de ativação tangente hiperbólica (*Tanh*) teve melhor desempenho, sendo capaz de generalizar o comportamento dos dados. Os registros com maiores MSE foram considerados *outliers* ou registros anômalos, e, dessa forma, segundo os autores, podem indicar indícios de fraudes em compras públicas. O uso de *Autoencoders* e outras redes neurais para classificação foi descartado para este trabalho, pois são técnicas que exigem grandes massas de dados para treinamento, e como será demostrado nos próximos capítulos, a quantidade de dados disponível foi considerada pequena.

O uso de técnicas de seleção de atributos e de preparação dos dados são apresentadas por Coussement [42], numa análise comparativa entre modelos de aprendizagem de máquina. O autor discorre sobre a relevância da preparação dos dados e como ela afeta a

performance do modelo de predição.

A pesquisa de Guyon *et al.* [46] desenvolve o método de seleção de variáveis utilizando SVM com RFE para seleção de variáveis de genes relacionados a câncer de cólon. Os autores demostram que a seleção realizada aumentou o poder de classificação dos casos de câncer, alcançando acurácia de 98%. O mesmo processo foi utilizado por Zhang *et al.* [81] para seleção de atributos mais importantes para detecção de fraudes e *fake-news* em vídeos e imagens.

Speck & Ferreira [17] atribuíram níveis de risco às modalidade de licitação através da análise de aspectos de procedimentos formais exigidos, grau de competitividade e discricionalidade do gestor na tomada de decisão. Segundo os autores, procedimentos mais formais, mais transparentes e com menor grau de discricionalidade tentem a ser menos suscetíveis à fraudes. Eles classificaram a dispensa de licitação e a inexibilidade com alto risco, a carta-convite e o concurso com médio risco e o leilão, o pregão, a concorrência e a tomada de preços com baixo risco.

A partir do trabalho de Speck & Ferreira [17], Rodidigues & Notato [16] verificaram a relação entre a modalidade de licitação e o risco de ocorrência de fraudes nos processos licitatórios realizados pelos municípios baianos. A partir de dados extraídos de 152 relatórios de controle interno da CGU, realizados entre 2004 e 2014, os autores aplicaram testes de hipóteses para verificar a relação entre a ocorrência de fraude e a modalidade aplicada. Eles concluíram que a carta convite e dispensa de licitação apresentaram maiores percentuais de irregularidades, enquanto que concorrência e pregão apresentaram os menores indicativos.

Conforme apresentado, foram encontradas publicações com aplicações de *machine learning* em dados de compras públicas com foco na modalidade pregão eletrônico. Desta forma, este trabalho difere dos demais quando aplica modelos de inteligência artificial observando o comportamento dos dados primeiramente em sua totalidade e depois segregados de acordo com a modalidade de licitação. Outro aspecto relevante da pesquisa é o estudo dos atributos que influenciam a ocorrência de irregularidades no conjunto total dos dados e compara o resultado com as compras agrupadas por modalidades.

Capítulo 3

Solução Proposta

A solução proposta consiste na junção de um módulo que extrai e processa dados das fontes primárias, outro que calcula o indicador de risco de irregularidades em compras e outro que disponibiliza a licitação em um ranking. A Figura 3.1 apresenta uma visão conceitual da solução proposta neste trabalho.

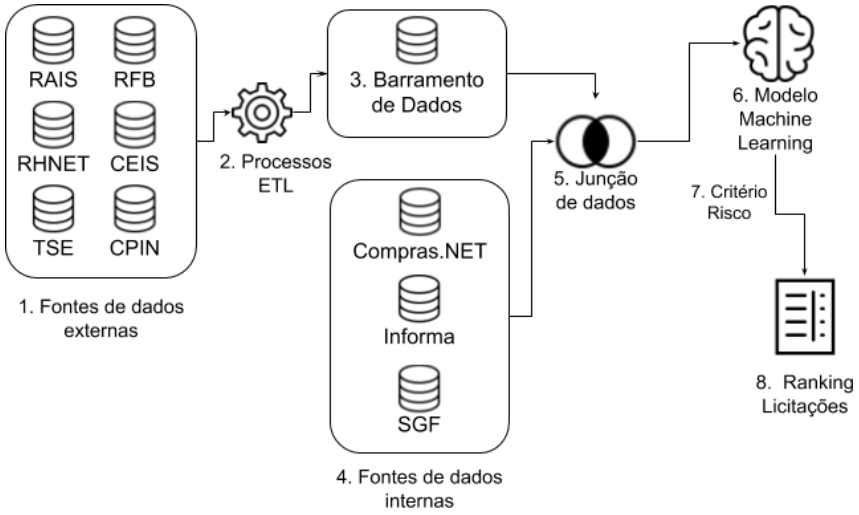

Figura 3.1: Solução Proposta.

Fonte: Próprio autor

As atividades apresentadas na Figura 3.1 são as seguintes:

1. Fonte de dados externa: são fontes de dados primárias externas ao Tribunal. Podem ser bases de dados abertas, tais como CEIS, bases de dados de órgãos sob jurisdição do TCE, como RH-NET e outras bases disponíveis através de acordos de cooperação com outros órgãos da administração pública. As bases de dados serão apresentadas na Seção 3.2.

2. Processos ETL: são processos de extração, transformação e limpeza de dados das fontes de dados externas, através de rotinas automatizadas que dependem do estudo de cada base de dados, entendimento dos atributos, tipo de dado e periodicidade de atualização. Neste projeto as rotinas foram escritas em linguagem *Python* e seu resultado foi armazenado em ambiente Big Data do TCE-GO.

3. Barramento de dados: ambiente Big Data do Tribunal, disponibilizado pela área de tecnologia da informação, contendo base de dados com dados das fontes externas processadas e validadas pelas rotinas de ETL.

4. Fontes de dados interna: são bases de dados de sistemas de gestão do Tribunal, disponíveis a qualquer tempo, sem necessidade de processos complexos de extração. As bases de dados serão apresentadas na Seção 3.2.

5. Junção de dados: rotinas automatizadas de cruzamento de dados para compor dados necessários para aplicação do modelo de *machine learning*.

6. Modelo de *machine learning*: modelo construído através de treinamento supervisionado com dados históricos de análise prévia de editais de licitações analisadas pela área de fiscalização do Tribunal. A resposta do modelo é o indicador de risco de irregularidades em licitações.

7. Critério de risco: indicador de indício de irregularidade em licitações calculado através de modelo de *machine learning*. Esse indicador varia entre zero e um.

8. Ranking de Licitações: lista ordenada de licitações que serve como instrumento de gestão e priorização para as áreas de fiscalização do Tribunal, para seleção de objetos de controle que permitam ação mais efetiva do controle externo prévio e concomitante no Estado de Goiás.

A construção do modelo de aprendizado de máquina da Etapa 6 da Figura 3.1 seguiu as fases do CRISP-DM. A Figura 3.2 apresenta uma visão conceitual da construção do modelo de aprendizado.

1. Entendimento do negócio: executa a etapa de entendimento do negócio, conforme descrito na Seção 3.1.

2. Entendimento dos Dados: a partir da necessidade do negócio apontada na Seção 3.1, realiza a coleta de dados a partir de licitações que foram previamente analisadas pelo Tribunal. Aplicam-se processos de limpeza nas bases de dados descritas na Seção 3.2.

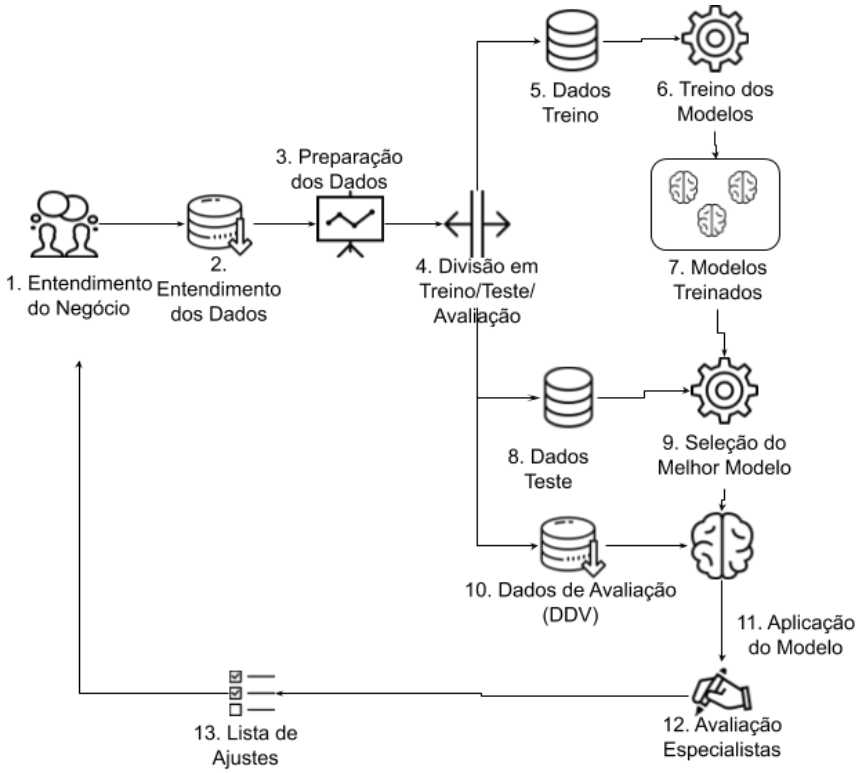

Figura 3.2: Modelo Mineração.

Fonte: Próprio autor

3. Preparação dos Dados: aplicam-se métodos estatísticos para entendimento dos dados. Os dados são padronizados. Aplicam-se métodos para seleção dos atributos mais relevantes. Esta etapa está detalhada na Seção 3.3

4. Divisão de dados em treino/teste e avaliação: os dados disponíveis são divididos em dados de treino/teste correspondentes a 90% dos dados totais.

5. Treino dos modelos: os modelos de aprendizado de máquina são treinados e testados com *cross-validation* com os dados de treino/teste. As métricas de avaliação são coletadas. Essa etapa está detalhada na Seção 3.4

6. Seleção do melhor modelo de classificação e melhor método de seleção de variáveis: avalia-se os modelos e métodos de seleção a partir das métricas coletadas. Aquele com melhor resultado é selecionado.

7. Aplicação do modelo: aplica-se o modelo aos dados das licitações não avaliadas para verificar a qualidade do treinamento.

8. Avaliação por especialistas: o resultado é posto para análise pelos especialistas em licitação do TCE-GO, que podem sugerir ajustes.

Na na Seção 2.1.2 foram apresentados os conceitos básicos sobre as licitações segundo as normas de Direito brasileiro. Cada modalidade possui particularidades com relação a prazos, forma de execução, forma de participação, regras de publicidade, composição de membros de gestão, dentre outras. Dessa forma, as etapas apesentadas na Figura 3.2 são realizadas considerando-se a fase (Edital/Disputa) e a modalidade. Para simplificar o processamento, foi solicitado apoio dos especialistas para agrupar modalidades com características mais semelhantes, conforme apresentado na Seção 2.1.2. Assim, as modalidades foram agrupadas para análise conjunta da seguinte forma:

- Todas Modalidades: análise total dos dados independentemente da modalidade
- Pregão: modalidades Pregão Presencial e Pregão Eletrônico
- Concorrência: modalidades Concorrência e Tomada de Preços
- Dispensa: as Dispensas de Licitação e Inexigibilidade Licitação

A Figura 3.3 apresenta de forma resumida o modo como os dados foram processados.

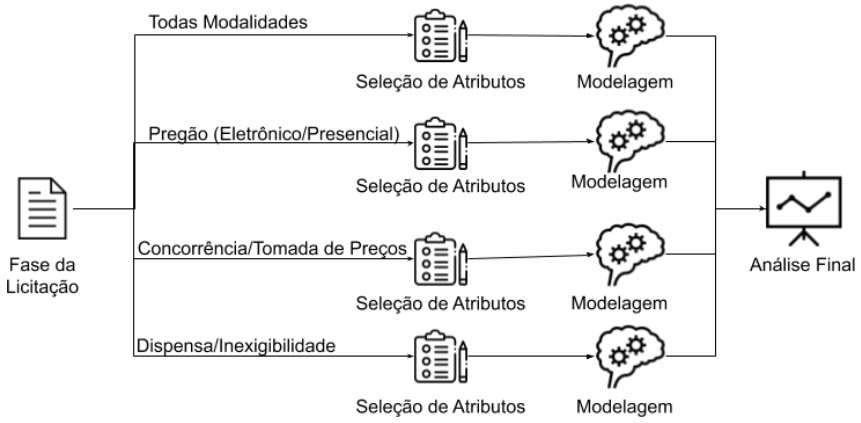

Figura 3.3: Processamento segundo fase e grupos por modalidade.

Fonte: Próprio autor

3.1 Entendimento do Negócio

Parte da etapa de entendimento do negócio foi apresentada ao longo de outras seções. A Seção 2.1 discorreu sobre o exercício do controle externo pelo Tribunal de Contas do

Estado de Goiás na fiscalização das compras públicas. Na Seção 2.1.2 foram apresentados os principais conceitos sobre licitações que são abordados neste trabalho e a Seção 2.1.3 abordou as principais irregularidades em compras públicas encontradas na literatura.

Os conjuntos de dados e informações foram agrupados conforme a fase da licitação:

1. Fase de Edital: dados que existem no momento da publicação do edital, termo de referência ou instrumento similar.
2. Fase de Disputa: dados que são criados durante a etapa de disputa ou seleção do fornecedor da licitação, dispensa ou inexigibilidade.

Para cada nova licitação publicada nos sistemas de compras, deve ser calculado o risco de irregularidades e o certame deve ser posto no *ranking* de editais para priorização pelos auditores do TCE-GO.

3.2 Entendimento dos Dados

Os dados utilizados neste trabalho são provenientes de diversas fontes das quais o Tribunal de Contas possui acesso direto, que são as Bases Internas. Ou acesso por acordos de cooperação com outros entes da administração pública, que são as Bases Externas. Nesta Seção é realizada a descrição de cada base de dados e de quais informações foram apontadas como relevantes por especialistas.

3.2.1 Bases de Dados Internas

Bases de Dados de Licitações

No Estado de Goiás, cada unidade administrativa é responsável por gerir suas aquisições e cada um possui suas próprias bases de dados de compras, como por exemplo, o Compras-NET do Poder Executivo.

No entanto, por força de Resolução do TCE-GO, todos os órgãos do Estado devem manter o Sistema Informa atualizado com os dados das suas licitações. Essa plataforma foi lançada pelo Tribunal em 2017 com o objetivo de centralizar os dados de aquisições do Estado de Goiás. Dessa forma, o Sistema Informa e o Compras-NET são as bases de dados de licitações para este trabalho. Os portais de transparência de alguns órgãos foram utilizados de forma complementar, como, por exemplo, o Portal de Licitações da Agência Goiana de Infraestrutura e Transportes (Goinfra) [1], o Portal de Transparência

[1]Disponível em: http://177.201.114.167/portal_licitacao/

do Poder Executivo[2], o Portal de Licitações da Secretaria Estadual da Saúde [3] e o Portal de Compras da Centrais de Abastecimento de Goiás (Ceasa-GO)[4].

Para este trabalho, foram levantadas as licitações publicadas entre janeiro de 2014 a dezembro de 2019, totalizando 32493 certames, com valor adjudicado de aproximadamente R$ 31,3 bilhões, conforme detalhado na Tabela 3.1.

Tabela 3.1: Aquisições públicas entre 2014 e 2019 no Estado de Goiás

Modalidade	**Qtd. Licitações Previstas**	**Valor Adjudicado**	**% Valor**
Pregão Eletrônico	13175	R$ 10.021.840.009,43	32,0%
Concorrência	1646	R$ 9.138.260.753,12	29,2%
Dispensa de Licitação	10939	R$ 5.876.955.109,83	18,8%
Pregão Presencial	855	R$ 4.218.092.455,95	13,5%
Inexibigilidade	4715	R$ 1.711.577.649,17	5,5%
Tomada de Preço	951	R$ 337.356.596,41	1,1%
Convite	212	R$ 16.277.285,96	0,1%
Total	**32493**	**R$ 31.320.359.859,87**	**100%**

A Tabela 3.1 evidencia que as licitações com modalidade Pregão Eletrônico representaram 32% do valor adjudicado, seguido da Concorrência com 29,2%. As modalidades Dispensa de Licitação e Inexibilidade representam juntas 24,3% do valor das aquisições, o que indica que o Estado de Goiás tem realizado sistematicamente aquisições por compra direta, e, conforme discutido na Seção 2.1.3, são modalidades que apresentam maior risco de irregularidade.

Atributos de Interesse

As bases de dados de licitação fornecem os atributos básicos para o treinamento dos modelos. Eles são os seguintes:

- Atributos da licitação
  - **lic_exclusiva_me:** indica se a licitação é exclusiva para micro ou pequena empresa.
  - **lic_lotes_vencidos:**: quantidade de lotes vencidos na licitação. Nas licitações de lances por item, essa variável contém o número de itens que podem ser vencidos, possibilitando capturar o grau de divisão que o certame teve e o grau de disputa da licitação.
  - **lic_menor_preco_vencedor:** indica se em todos os lotes/itens o menor preço ofertado foi o vencedor.

[2]Disponível em: http://www.transparencia.go.gov.br/portaldatransparencia/
[3]Disponível em https://www.saude.go.gov.br
[4]Disponível em: https://www.ceasa.go.gov.br/acesso-a-informacao/

- **lic_prop_desagio_lotes_vencidos:** variável criada neste trabalho que armazena o deságio na licitação. O deságio é a relação entre a diferença do valor estimado e o valor adjudicado por lote vencido.
- **lic_prop_lotes_vencidos:** variável criada neste trabalho que armazena a relação entre os lotes/itens estimados e aqueles que foram vencidos.
- **lic_prop_participantes_desclassificacoes:** variável criada neste trabalho que armazena a relação entre o total de participantes da licitação e o total de participantes desclassificados.
- **lic_prop_participantes_por_lote:** variável criada neste trabalho que armazena a relação entre total de participantes e o total de lotes.
- **lic_prop_participantes_recursos:** variável criada neste trabalho que armazena a relação entre participantes e a quantidade de recursos apresentados na licitação.
- **lic_prop_participantes_valor_estimado:** variável neste trabalho que armazena a relação entre participantes da licitação e o valor estimado.
- **lic_qtd_desclass:** quantidade de desclassificações/inabilitações na etapa de disputa da licitação.
- **lic_qtd_itens:** quantidade de itens da licitação. Nas modalidades concorrência e tomada de preços para obras ou serviços de engenharia, esta variável armazena a quantidade de itens de maior relevância técnica e de valor significativo, conforme art. 30 da Lei 8.666/93.
- **lic_qtd_lotes:**: quantidade de lotes da licitação.
- **lic_qtd_participantes:** quantidade total de participantes da licitação. Nos casos de Dispensa ou Inexigibilidade, foram consideradas as empresas consultadas na etapa de orçamentação feita pelo órgão contratante.
- **lic_qtd_recursos:** quantidade de recursos apresentados pelos licitantes.
- **lic_tipo_licitacao:** tipo da licitação.
- **lic_valor_adjudicado:** valor total em reais adjudicado na licitação.
- **lic_valor_estimado:** valor estimado em reais na licitação.

- Atributos do órgão contratante
  - **orgao_qtd_licitacoes:** quantidade de licitações do órgão no ano da licitação.
  - **orgao_qtd_aditivos:** quantidade de aditivos em contratos que o órgão fez no ano da licitação.

- **orgao_qtd_licitacoes_dispensa:** quantidade de dispensas de licitação que o órgão realizou no ano da licitação.
- **orgao_qtd_licitacoes_inexigibilidade:** quantidade de inexigibilidades de licitação que o órgão realizou no ano da licitação.
- **orgao_qtd_licitacoes_pregao:** quantidade de licitações com modalidade "pregão" que o órgão realizou no ano da licitação.
- **orgao_qtd_licitacoes_convite:** quantidade de licitações com modalidade "convite" que o órgão realizou no ano da licitação.
- **orgao_qtd_licitacoes_concorrencia:** quantidade de licitações com modalidade "concorrência" que o órgão realizou no ano da licitação.

- Atributos para validação e atributos chaves para outras bases de dados
  - **val_orgao_cnpj:** número do CNPJ do órgão contrante.
  - **val_cpf_presidente:** número do CPF que pode ser do presidente da comissão de licitação para modalidades Concorrência, Tomada de Preço, Convite e casos de compra direta ou o CPF do pregoeiro para Pregão Presencial ou Pegão Eletrônico.
  - **val_cnpj_participantes:** lista de CNPJ dos participantes da licitação com indicação de quais foram os vencedores.
  - **val_situacao_licitacao:** indica se a licitação foi finalizada, está em andamento, se foi cancelada ou revogada.

Bases de Dados de Análises das Áreas Técnicas do TCE-GO

As análises prévias de licitação realizadas no âmbito do TCE-GO se tornam processos no Sistema de Processo Eletrônico do Tribunal, plataforma estruturada chamada e-TCE [5]. Subsidiariamente, as áreas técnicas alimentam planilhas eletrônicas com dados das inconsistências encontradas nas análises e informações complementares.

Foi realizada pesquisa de processos no sistema e-TCE filtrando o assunto do processo como "Licitação"e data da realização entre 2014 e 2019, retornando um total de 1637 processos. Desses, foram eliminados aquelas que não possuíam nenhum documento do tipo "Instrução Técnica", que é a peça processual que contém a manifestação dos auditores a respeito do procedimento licitatório. Assim restaram 566 processos de licitações nesta amostra.

[5]Sistema e-TCE: https://etce.tce.go.gov.br

Foi feito cruzamento dos processos selecionados no e-TCE com a base de dados de análises processuais das áreas técnicas. Além disso, cada processo foi aberto no sistema e foi feita leitura da "Instrução Técnica". A análise técnica pode ter como encaminhamento:

1. Devolução sem análise de mérito: o auditor emite opinião de que a análise não deve ser feita por diversos motivos, tais como, processo perdeu objeto ou a licitação foi cancelada ou revogada. Nestes casos, não há apontamentos sobre a ausência ou presença de irregularidades e o processo segue para arquivamento sem análise do mérito.

2. Encaminhamento para diligência: o auditor precisa de informações adicionais para emitir sua opinião sobre o processo licitatório e as solicita ao órgão que realizou o certame. Esses processos não podem ser utilizados neste trabalho, pois ainda não há apontamentos sobre a ausência ou a presença de irregularidades.

3. Emissão da instrução técnica conclusiva com análise de mérito: o auditor analisa todo o processo e emite opinião pela legalidade da licitação ou pela sua irregularidade. Apenas esses processos podem ser utilizados neste trabalho.

Após filtrar os casos de processos que continham instrução técnica conclusiva com análise de mérito, restaram 466 licitações na amostra. Dessas, foi feito cruzamento com as bases de dados de licitações do Sistema Informa e Sistema Compras-NET, restando 369 certames. As licitações não localizadas nos dois sistemas podem ser de órgãos que não utilizam o Compras-NET, como por exemplo, Indústria Química do Estado de Goiás (Iquego), Ceasa-GO, Companhia Saneamento de Goiás S/A (Saneago) e licitaram antes de 2017, data de entrada no Sistema Informa, ou não enviaram dados pelo Sistema Informa, contrariando Resolução do TCE-GO. Nesses casos, não há como obter informações estruturadas das compras realizadas. A Tabela 3.2 apresenta um resumo dos quantitativos de licitações que foram capturadas.

Tabela 3.2: Visão geral sobre os dados coletados

Descrição	**Total**	**% Sobre Total**
Total de licitações no período	32493	100,0%
Total de processos no e-TCE	1637	5,0%
Total de licitações com análise de mérito	466	1,4%
Total de licitações localizadas nos sistemas de compras	369	1,1%

A Tabela 3.2 demostra que a amostra disponível para modelagem é pequena em relação ao total de licitações que existem. No entanto, ela representa aproximadamente 80% de todas licitações que tiveram análise de mérito pelo TCE-GO no período de estudo.

A Figura 3.4 apresenta a distribuição do conjunto de dados por modalidade. A maior parte se refere à modalidade Pregão Eletrônico, seguido por Concorrência.

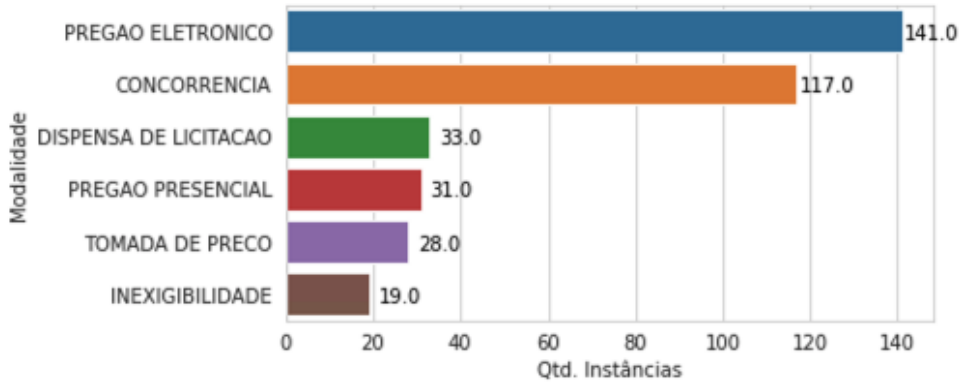

Figura 3.4: Licitações por modalidade.

Fonte: Próprio autor

A amostra contém licitações feitas por 29 órgãos da Administração Pública Estadual. A maior quantidade, 39%, se concentra na Agência Goiana de Infraestrutura e Transportes, atual Goinfra. Isso se deve ao fato de que a Goinfra é órgão estadual responsável por obras públicas, devido aos altos valores envolvidos e a complexidade técnica, o TCE-GO possui uma área especializada em análise de obras e serviços de engenharia. A Figura 3.5 apresenta a seis primeiras unidades administrativas com mais licitações na amostra.

A Figura 3.6 apresenta a distribuição dos processos de compras por ano de realização. Na maioria dos casos, a análise do Tribunal que contém a instrução técnica conclusiva ocorreu após o término da licitação. As exceções são quando o certame contém grave afronta a alguma norma ou visível prejuízo ao erário, situações em que o TCE-GO atua para impedir o andamento da licitação através de medidas cautelares. Assim, na Figura 3.6 observa-se que o ano de 2014 contém a maior quantidade de processos com decisão, pois a quantidade de licitações com decisões da área técnica por ano se acumula com o passar do tempo.

A Figura 3.7 apresenta o balanceamento da amostra comparando os resultados entre licitações com irregularidades e sem irregularidades. Percebe-se que não há um grande desbalanceamento entre os dados, mas chama atenção o fato de que a quantidade de licitações com irregularidades é maior que a classe das sem irregularidades. Isso pode indicar que os critérios adotados atualmente pelo TCE-GO na seleção de licitações para fins de auditoria não estão otimizados para concentrar o esforço naquelas com maior risco. Assim, pelo princípio da eficiência, seria mais adequado que os limitados recursos disponíveis fossem concentrados nos cenários mais críticos de atuação.

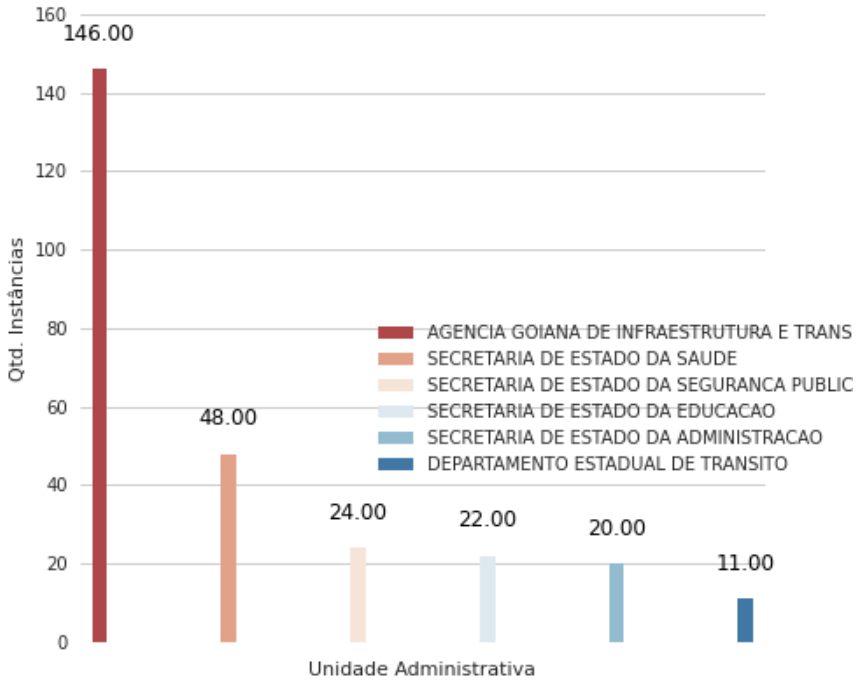

Figura 3.5: Licitações por unidade administrativa.

Fonte: Próprio autor

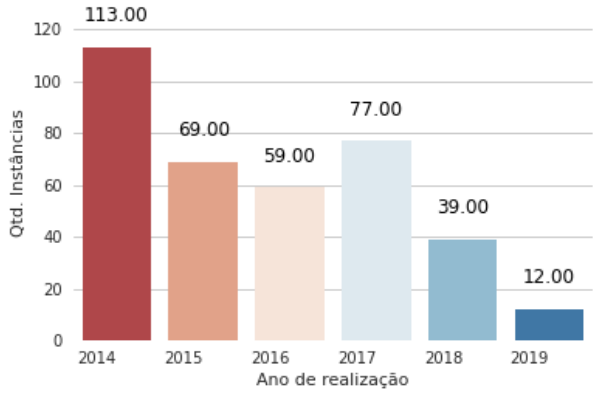

Figura 3.6: Distribuição dos processos de compras por ano de realização.

Fonte: Próprio autor

Na amostra foram encontradas 353 ocorrências de tipologias de irregularidades, apresentadas na Seção 2.1.3, distribuídas em 191 licitações. A Tabela 3.3 apresenta a quantidade de ocorrência das tipologias.

Conforme a Tabela 3.3, foram encontradas 22 tipologias diferentes nas licitações analisadas. No entanto, 54% das 353 ocorrências se concentram em apenas quatro tipologias:

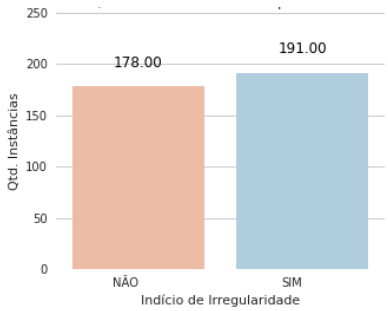

Figura 3.7: Balanceamento do atributo "Indício de Irregularidade"

Fonte: Próprio autor

Tabela 3.3: Tipologias encontradas na amostra.

Tipologia	**Qtd. Ocorrências**	**%**
Restrição de competitividade - Cláusulas restritivas	61	17,3%
Superfaturamento ou sobrepreço em orçamentos	53	15,0%
Ausência de projeto básico e orçamentos em planilhas	39	11,0%
Ausência ou deficiência de pesquisa de preços	39	11,0%
Outras irregularidades	33	9,3%
Ausência de programação de desembolso financeiro	23	6,5%
Ausência de declaração do ordenador de despesa	11	3,1%
Objeto impreciso, genérico, incompressível ou incompleto	11	3,1%
Utilização de critérios de habilitação ou julgamento sem previsão em edital	10	2,8%
Ausência de orçamento detalhados e preços unitários	9	2,5%
Uso de modalidade indevida	8	2,3%
Ausência ou deficiência na publicidade e transparência	7	2,0%
Composição da comissão de licitação irregular	6	1,7%
Conluio	6	1,7%
Fraude em licitação	6	1,7%
Desclassificação/Inabilitação indevida de licitante	5	1,4%
Inobservância da previsão de preferência a contratações de ME e EPP	5	1,4%
Uso indevido de Pregão Presencial em detrimento de Pregão Eletrônico	5	1,4%
Ausência de comprovação de capacidade técnica	4	1,1%
Ausência ou aplicação incorreta de BDI	4	1,1%
Ausência ou deficiência no de planejamento prévio da contratação	4	1,1%
Dispensa ou inexigibilidade sem fundamentação legal	4	1,1%
Total	**353**	-

(1) “Restrição de competitividade - Cláusulas restritivas”, (2) “Superfaturamento ou sobrepreço em orçamentos”, (3) “Ausência de projeto básico e orçamentos em planilhas” e (4) “Ausência ou deficiência de pesquisa de preços”. Todas essas se relacionam à confecção ilegal do instrumento convocatório na fase interna da licitação pelos os agentes públicos. Além disso, essas tipologias de irregularidades, se não tratadas, podem gerar dano ao erário.

Atributos de Interesse

Essa base de dados possui os seguintes atributos de interesse:

- Atributos da licitação
  - **lic_indicio_irregularidade:** indica se as áreas de fiscalização do TCE-GO encontraram uma ou mais irregularidades da Tabela 2.1 da Seção 2.1.3.
- Atributos de validação e atributos chaves para outras bases de dados
  - **val_desc_tipologia_indicio_irregularidade:** descrição do indício de irregularidade apontado pela área técnica, conforme Tabela 2.1 da Seção 2.1.3.
  - **val_numeroprocesso_tce:** número do processo no sistema e-TCE.
  - **val_numeroprocesso_jurisdicionado:** número do processo no sistema de compras do órgão jurisdicionado.

Bases de Sanções do TCE-GO

O TCE-GO tem prerrogativa de julgar contas de gestores enviadas anualmente ao Tribunal. Os gestores com contas rejeitadas por irregularidade insanável, em decisão irrecorrível, são mantidos em um cadastro disponível para a sociedade[6].

Além disso, o Tribunal mantém internamente o cadastro de pessoas físicas ou jurídicas que sofreram sanções por má gestão de recursos públicos ou foram responsabilizados por fraudes, irregularidades ou ação ilegal em atos administrativos identificados em ações de controle.

Atributos de Interesse

Esta base de dados possui os seguintes atributos de interesse:

- Pessoas com poder de decisão sobre a licitação
  - **gestor_qtd_sancoes_tce:** quantidade de punições ao gestor.
  - **presidente_qtd_sancoes_tce:** quantidade de punições ao presidente da comissão de licitação ou pregoeiro.
- Participantes
  - **vencedor_qtd_sancoes_tce:** quantidade de punições aplicadas ao vencedor.

[6]Disponível em: https://portal.tce.go.gov.br/contas-irregulares

- **vencedor_valor_sancoes_tce:** valor total das punições aplicadas ao vencedor.
- **lic_qtd_sancoes_part_tce:** soma da quantidade de punições aplicadas aos participantes.

Bases de Informações Estratégicas do TCE-GO

O TCE-GO, através da unidade de inteligência chamada "Serviço de Informações Estrategicas", possui acesso a bases de dados com informações estratégicas de pessoas físicas e jurídicas alcançadas pela jurisdição do Tribunal. Essa base contém 13 atributos considerados relevantes para esse estudo, segundo opinião de especialistas. No entanto, essas informações têm caráter sigiloso e seu conteúdo não pôde ser apresentado neste trabalho. Assim, essas variáveis foram renomadas para garantir o sigilo da informação.

Atributos de Interesse

Esta base de dados possui os seguintes atributos de interesse:

- Pessoas com poder de decisão sobre a licitação
  - **gestor_sie_v1**
  - **gestor_sie_v2**
  - **gestor_sie_v3**
  - **gestor_sie_v4**
  - **gestor_sie_v5**
  - **gestor_sie_v6**
  - **gestor_sie_v7**
- Participantes
  - **vencedor_sie_v1**
  - **vencedor_sie_v2**
  - **vencedor_sie_v3**
  - **vencedor_sie_v4**
  - **vencedor_sie_v5**
  - **vencedor_sie_v6**
  - **vencedor_sie_v7**
  - **vencedor_sie_v8**

Sistema de Gestão de Fiscalização - SGF

O Sistema de Gestão de Fiscalização do TCE-GO (SGF) forma a base de conhecimento de propriedade do TCE-GO que contém informações históricas das fiscalizações realizadas pelo Tribunal. Trata-se de sistema implantando em 2012 e amplamente utilizado nas fases construção dos papéis de trabalho das auditorias. Cada ação de fiscalização é tratada como um projeto que segue o ciclo *Plan, Do, Check, Action* (PDCA).

Atributos de Interesse

Esta base de dados possui os seguintes atributos de interesse:

- Atributos das fiscalizações
  - **orgao_total_fiscalizacoes_tce_ano:** armazena a quantidade de fiscalizações que o TCE-GO realizou em órgão estadual por ano.
  - **orgao_total_fiscalizacoes_tce:** armazena a quantidade de fiscalizações que o TCE-GO realizou em órgão estadual em todo período de levantamento de dados deste trabalho.

Sistema Rol de Responsáveis

O Sistema Rol de Responsáveis é uma base de dados mantida pelo TCE-GO que contém informações sobre os ordenadores de despesa no Estado de Goiás. Nessa plataforma é possível coletar informações sobre as pessoas com poder de decisão em determinado órgão estadual na data de publicação de uma licitação.

Atributos de Interesse

Esta base de dados possui os seguintes atributos de interesse:

- Gestor do órgão
  - **val_cpf_gestor:** armazena o CPF do gestor do órgão na data da licitação.

Sistema de Gestão Financeira e Orçamentária do Estado de Goiás

O Sistema de Execução Financeira e Orçamentária do Estado de Goiás (SIOFI-NET) é ferramenta oficial de gestão financeira e contém os pagamentos realizados aos fornecedores. Logo, trata-se de uma base rica em informações sobre o planejamento orçamentário e financeiro do órgão contratante, do volume financeiro gasto para aquisições de bens, produtos ou serviços, pagamentos aos participantes da licitação e quantidade de contratos que esses participantes possuem com o Estado.

Atributos de Interesse

Esta base de dados possui os seguintes atributos de interesse:

- Órgão contratante
  - **orgao_total_dotacao_orcado:** total do orçamento previsto para órgão contratante no ano da licitação.
  - **orgao_total_dotacao_cred_especial:** total de créditos especiais no orçamento do órgão contratante no ano da licitação.
  - **orgao_total_dotacao_cred_extraordinario:** total de créditos extraordinários no orçamento do órgão contratante no ano da licitação.
  - **orgao_total_dotacao_autorizado:** total da dotação orçamentária autorizada para o órgão contratante no ano da licitação.
  - **orgao_total_saldo_empenho:** total empenhado pelo órgão contratante em naturezas de despesas para aquisição de bens, produtos ou serviços no ano da licitação.
  - **orgao_total_saldo_liquidado:** total liquidado pelo órgão contratante em naturezas de despesas para aquisição de bens, produtos ou serviços no ano da licitação.
  - **orgao_total_saldo_pago:** total pago pelo órgão contratante em naturezas de despesas para aquisição de bens, produtos ou serviços no ano da licitação.
- Participantes da licitação
  - **vencedor_total_pago_goias:** total pago ao vencedor da licitação entre 2014 e 2019.
  - **vencedor_qtd_orgaos_pagadores_goias:** quantidade de órgãos que efetuaram pagamentos para o vencedor da licitação entre 2014 e 2019.

3.2.2 Bases de Dados Externas

São bases de dados de terceiros que o TCE-GO tem acesso por força de lei ou através de acordos de cooperação com outros entes da administração pública.

Sistema RH-NET e Base de Dados de Servidores Públicos Estaduais

O Poder Executivo mantém uma base de dados informações cadastrais de todos os servidores ativos, inativos, pensionistas e temporários, com dados desde 2003. Nessa plataforma

existem informações sobre cargos ocupados, tipo de vínculo (efetivo, comissionado, temporário), remuneração, data de ingresso no serviço público. Além disso, o Serviço de Informações Estratégicas mantém dados de todos os servidores públicos estaduais, ativos e inativos, dos poderes Legislativo, Judiciário, do Ministério Público Estadual, Defensoria Pública, Tribunal de Contas dos Municípios e empregados públicos de empresas públicas do Estado.

Essas bases de dados, em conjunto com o Sistema Rol de Responsáveis, permitem identificar pessoas que podem estar em situação de conflito interesse na condução de um processo licitatório, quando, por exemplo, existem vínculos entre gestores do órgão ou membros da comissão de licitação com sócios ou representantes das empresas participantes. Também permite verificar possível influência do tipo do vínculo do servidor que tem poder de decisão na licitação.

Atributos de Interesse

Esta base de dados possui os seguintes atributos de interesse:

- Pessoas com poder de decisão sobre a licitação
  - **gestor_grupo_cargo:**: tipo de vínculo do gestor do órgão contratante na data da licitação.
  - **presidentecomissao_grupo_cargo:** tipo de vínculo do presidente ou pregoeiro na data da licitação.
- Participantes da licitação
  - **vencedor_qtd_socios_servidores:**: quantidade de sócios da empresa vencedora presentes nas folhas de pagamentos do Estado nos 5 últimos anos antes da data da licitação.
  - **participantes_qtd_socios_servidores:** quantidade de sócios das empresas participantes presentes na folha de pagamentos do Estado nos 5 últimos anos antes da data da licitação.

Cadastro de Pessoas Físicas e Jurídicas

A Receita Federal do Brasil mantém base de dados com informações de todas as empresas, chamada de Cadastro Nacional de Pessoas Jurídicas. Essa base contém informações sobre porte da empresa, endereço, contadores, quadro societário e atividade econômica. De forma semelhante, a Receita Federal detém o Cadastro Nacional de Pessoas Físicas.

Através do quadro societário, é possível avaliar se duas ou mais empresas com sócios em comum participam da mesma licitação, o que pode violar o princípio da competitividade.

Outra base de interesse do controle externo é o Cadastro Nacional de Atividades Econômicas, que categoriza as empresas por atividade principal e secundárias. Essa informação é importante para, por exemplo, analisar se uma determinada empresa participa de uma licitação para um objeto que não é compatível com as suas atividades.

Atributos de Interesse

Esta base de dados possui os seguintes atributos de interesse:

- Pessoas com poder de decisão sobre a licitação
  - **gestor_qtd_empresas_periodo:** quantidade de empresas que o gestor do órgão foi sócio entre 2014 e 2019.
  - **presidentecomissao_qtd_empresas_periodo:** quantidade de empresas que o presidente da comissão de licitação ou o pregoeiro foi sócio entre 2014 e 2019.
- Participantes
  - **vencedor_sit_cadastral:** situação cadastral da empresa vencedora na data da licitação.
  - **vencedor_porte:** porte da empresa vencedora.
  - **vencedor_capital_social:** valor do capital social da empresa vencedora.
  - **vencedor_prazo_abertura_ate_licitacao:** prazo em dias entre a data de abertura da empresa vencedora até a data da licitação.
  - **vencedor_qtd_socios:** quantidade de sócios da empresa vencedora na data da licitação.
  - **vencedor_qtd_cnae:** quantidade de atividades registradas na Classificação Nacional de Atividades Econômicas (CNAE) da empresa vencedora na data da licitação.
  - **lic_qtd_partc_epp:** quantidade de participantes com porte Empresa de Pequeno Porte (EPP).
  - **lic_qtd_partc_me:** quantidade de participantes com porte Microempresa (ME).
  - **lic_qtd_partc_out_portes:** quantidade de participantes com porte diferente de EPP ou ME

Relação Anual de Informações Sociais (RAIS)

A RAIS é base de dados com informações sobre empregos formais registrado junto ao Ministério do Trabalho. Por força de lei, todas as empresas são obrigadas a informar anualmente dados dos seus empregados, como nome, CPF, cargo, data de entrada e saída, nível de escolaridade e remuneração.

Com essa base de dados é possível identificar vínculo entre empresas participantes de uma licitação (funcionários em comum) ou ainda avaliar o risco ao comparar a força de trabalho de uma empresa com aquela necessária para executar uma determinada atividade, como por exemplo, uma empresa com quatro funcionários cadastrados pleiteando um contrato de prestação de serviço de limpeza que requer 500 pessoas. Quando uma determinada empresa possui zero empregados pode-se atribuir a ela certo grau de risco de ser empresa de fachada.

Atributos de Interesse

Esta base de dados possui os seguintes atributos de interesse:

- Pessoas com poder de decisão sobre a licitação
  - **lic_qtd_pessoas_rais_participantes:** indica quantas vezes cada pessoa com poder de decisão (gestor, presidente de comissão e pregoeiro) teve vínculos empregatícios com participantes da licitação entre 2014 e 2019.
- Participantes
  - **vencedor_qtd_vinculos_empregaticios:** quantidade de empregados vinculados à empresa vencedora no ano da licitação.
  - **participantes_qtd_funcionarios_comum:** quantidade de empregados em comum entre as participantes do certame entre 2014 e 2019.

Bases da Justiça Eleitoral

As bases de dados do TSE permitem verificar se existem vinculação política entre os envolvidos no processo licitatório, tais como, se o sócio da empresa vencedora realiza doações a partido político, ou se o pregoeiro tem filiação partidária. Em ambos os casos, pode existir influência política na atuação desses agentes, o que acarreta risco ao processo de compra pública. Para simplificar o modelo, foram contados todos os vínculos políticos que determinada pessoa pode ter: candidato; filiado a partido; dirigente de partido.

Atributos de Interesse

Esta base de dados possui os seguintes atributos de interesse:

- Pessoas com poder de decisão sobre a licitação
  - **gestor_qtd_vin_pol_partidario:** quantidade de vínculos político-partidários que o gestor possui entre 2014 e 2019.
  - **pres_qtd_vinc_pol_partidario:** quantidade de vínculos político-partidários que o presidente da comissão de licitação ou pregoeiro possuem entre 2014 e 2019.
- Participantes da licitação
  - **vencedor_qtd_socios_vinc_pol_partidario:** soma da quantidade de vínculos político-partidários dos sócios da empresa vencedora.
  - **participantes_qtd_socios_vinc_pol_partidario:** soma da quantidade de vínculos político-partidários dos sócios das participantes.

Cadastro Nacional das Empresas Inidôneas - CEIS

O Cadastro Nacional das Empresas Inidôneas CEIS [7] é uma base de dados que contém relação de empresas que possuem restrição de participar de licitações ou contratar com Poder Público.

Esse cadastro é realizado por toda a administração pública, em todas as esferas, conforme obrigação prevista no artigo 23 da Lei 12.846 (Lei Anti-corrupção) [8]. O CEIS está disponível para download no portal de transparência do Governo Federal [9].

Atributos de Interesse

Esta base de dados possui os seguintes atributos de interesse:

- Pessoas com poder de decisão sobre a licitação
  - **gestor_qtd_ocorrencia_ceis:** quantidade de punições aplicadas ao gestor.
  - **presidente_qtd_ocorrencia_ceis:** quantidade de punições aplicadas ao presidente de comissão ou ao pregoeiro.
- Participantes da licitação

[7] http://www.cgu.gov.br/assuntos/responsabilizacao-de-empresas/ sistema-integrado-de- registro-do-ceis-cnep

[8] http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2013/lei/l12846.htm

[9] http://www.portaltransparencia.gov.br

- **vencedor_qtd_ocorrencia_ceis:** quantidade de punições aplicadas a empresa vencedora no CEIS.
- **lic_qtd_ocorrencia_participantes_ceis:** soma da quantidade de punições aplicadas a empresas participantes no CEIS.

Cadastro Nacional das Empresas Punidas - CNEP

O Cadastro Nacional das Empresas Punidas é amparado pela mesma base legal que suporta a CEIS e contém as empresas que foram punidas pela administração pública. Essa informação pode ser relevante, pois um fornecedor que tenha sido punido repetidas vezes em uma unidade da federação pode cometer a mesma infração no Estado de Goiás.

Atributos de Interesse

Esta base de dados possui os seguintes atributos de interesse:

- Participantes
  - **vencedor_qtd_ocorr_cnep:** quantidade de punições da empresa vencedora no CNEP.
  - **lic_qtd_ocorr_partic_cnep:** soma da quantidade de punições das empresas participantes no CNEP.

Entidades Privadas sem Fins Lucrativos Impedidas (CEPIM)

O Cadastro de Entidades Privadas sem Fins Lucrativos Impedidas (CEPIM)[10], da mesma forma que o CEIS e CNEP, é mantido pela CGU e contem relação de empresas sem fins lucrativos impedidas de contratar com a administração pública por terem cometido irregularidades em outros contratos ou convênios.

Atributos de Interesse

Esta base de dados possui os seguintes atributos de interesse:

- Participantes
  - **vencedor_qtd_ocorr_cnep:** quantidade de punições aplicadas a empresa vencedora no Cadastro de Entidades Privadas Sem Fins Lucrativos Impedidas (CEPIM).
  - **lic_qtd_ocorr_partic_cnep:** soma da quantidade de punições aplicadas as empresas participantes no CEPIM.

[10]http://www.portaltransparencia.gov.br/ pagina-interna/603243-cepim

3.3 Preparação dos Dados

A Seção 3.2 descreve as bases de dados e os atributos de interesse deste trabalho. A etapa de coleta de dados consiste na construção de rotinas que buscam dados nas fontes primárias apresentadas nas Seções 3.2.1 e 3.2.2 para persistir em uma base local.

Após análise cuidadosa de cada fonte de dados na Seção 3.2, foi constatado que cada uma possui diferentes sistemas gerenciadores de bancos de dados, diferentes meios de acessos (conexão direta via cliente, *Web-Service*, carga de arquivos CSV) e ainda diferentes periodicidades de carga (bases de licitações no Compras-NET são atualizadas diariamente, bases de folha de pessoal no RH-NET atualiza mensalmente e a RAIS possui atualização anual). Portanto, devido a complexidade envolvida na extração e processamento dos dados, foi necessário introduzir métodos de seleção de atributos, de modo a verificar quais dados e quais bases são mais relevantes.

A Tabela 3.4 apresenta um resumo das licitações coletadas.

Tabela 3.4: Totais de dados das licitações

Atributo	**Total**
Qtd. licitações	369
Qtd. licitações com irregularidades	191
Qtd. licitações sem irregularidades	178
Qtd. Órgãos (Jurisdicionados)	29
Qtd. Gestores dos órgãos	64
Qtd. Presidente/Pregoeiros	95
Qtd. Empresas Participantes	1164

O conjunto de dados é composto por 369 licitações, com 87 atributos coletados de diversas fontes, sendo o atributo "lic_indicio_irregularidade" a variável de interesse. Ela assume os valores "0" para "licitação sem irregularidades" ou "1" para licitações irregulares. Das 369 licitações, 344 foram finalizadas e adjudicadas e outras 25 não consagraram um vencedor, pois foram revogadas ou desertas. No entanto, como tiveram um instrução técnica conclusiva com análise de mérito emitida pelo TCE-GO sobre sua legalidade, foram utilizadas neste trabalho para a Fase de Edital. Nessa fase, os atributos relacionados à Fase de Disputa são removidos, restando 41 atributos.

Entre 2014 e 2019, as licitações em análise foram realizadas por 29 órgãos diferentes, com 64 gestores (Secretários, Presidentes) e 95 presidentes de comissão de licitação ou pregoeiros, com participação de 1164 empresas. Portanto, percebe-se a rotatividade das pessoas com poder de decisão nas licitações e nas empresas que competem pela melhor proposta.

As 344 licitações finalizadas foram utilizadas na Fase de Disputa. Uma licitação nessa fase pode ser vencida por mais de um fornecedor. Assim, decidiu-se intencionalmente

adicionar as informações dos vencedores no conjunto de dados assumindo-se a duplicação de algumas outras informações. Dessa forma, o conjunto completo para a Fase de Disputa contou com 548 instâncias e 87 atributos. Para essa fase, a Figura 3.8 apresenta a distribuição do conjunto de dados por modalidade. A maior parte se refere à modalidade Pregão Eletrônico, seguido por Concorrência.

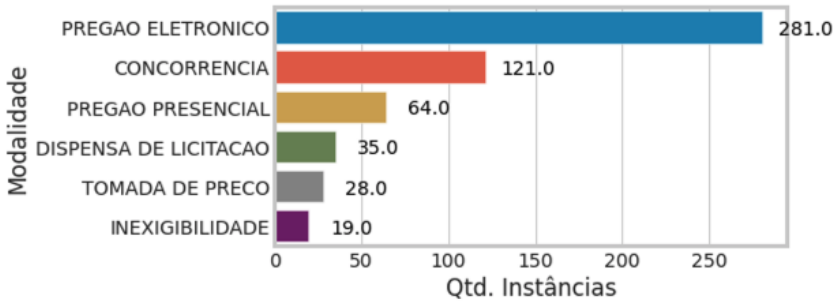

Figura 3.8: Licitações por modalidade na Fase de Disputa.

Fonte: Próprio autor

Extração e Limpeza de Dados

Os mais de 1.600 processos da Tabela 3.2 que contém análise de licitações cadastradas no Tribunal foram extraídos do sistema e-TCE e planilhados. Até o ano de 2019, antes de ajustes no sistema e-TCE e no Sistema SGF, os dados necessários para este trabalho estavam apenas nas peças processuais. Dessa forma, foi realizada a leitura das instruções técnicas conclusivas e anotadas as irregularidades identificadas pelos auditores. Foi também anotado o número de processo de aquisição do órgão que fez a licitação para cruzamento de dados da licitação no sistema do jurisdicionado. Essa etapa consumiu cinco meses de esforço deste trabalho.

As demais informações complementares foram extraídas das bases de dados através de consultas *SQL*, utilizando os parâmetros anotados nos processos cadastrados no Tribunal. Após a extração dos dados, foi realizada análise de cada um dos atributos verificando máximos, mínimos, repetição de valores, valores faltantes, variância e correlação. A Tabela 3.5 apresenta análise realizada para alguns atributos da base de dados.

Padronização e Variáveis *Dummy*

Conforme descrito na Seção 2.4.3, foi realizada a padronização dos dados com tipo numérico com o algoritmo *StandardScaler* do pacote *sklearn.preprocessing*.

Tabela 3.5: Atributos numéricos na base de dados.

Atributo	Total	Média	Valor Min.	Valor Max.	Coeficiente de Variação
lic_valor_estimado	R$ 7.543.898.336,84	R$ 20.444.168,93	R$ 5,00	R$ 1.292.154.958,00	4,1
lic_valor_adjudicado	R$ 5.424.788.059,91	R$ 14.812.144,05	R$ 0	R$ 987.045.175,30	4,6
lic_qtd_participantes	1164,00	6,33	1,00	46,00	0,92
lic_qtd_recursos	253,00	0,69	0,00	19,00	2,60
lic_qtd_desclass	528,00	1,43	0,00	68,00	3,39
lic_qtd_lotes	1394,00	3,78	1,00	103,00	2,22
lic_qtd_itens	4731,00	12,82	1,00	320,00	2,54

Para a Fase de Edital, as variáveis categóricas da lista abaixo foram transformadas em variáveis *Dummy*, adicionando 10 atributos ao conjunto de dados, que passou a contar com 369 instâncias e 51 atributos.

- gestor_grupo_cargo: 3 valores distintos
- pres_grupo_cargo: 4 valores distintos
- lic_exclusiva_me: 2 valores distintos
- lic_tipo_licitacao: 5 valores distintos

De modo semelhante, as variáveis categóricas aplicáveis à Fase de Disputa foram transformadas em *Dummy*, aumentando o número de atributos em 15, de forma que o conjunto de dados passou a ter 548 instâncias e 102 atributos.

- gestor_grupo_cargo: 3 valores distintos
- pres_grupo_cargo: 4 valores distintos
- lic_exclusiva_me: 2 valores distintos
- lic_tipo_licitacao: 5 valores distintos
- vencedor_sit_cadastral: 3 valores distintos
- vencedor_porte: 3 valores distintos
- lic_menor_preco_vencedor: 2 valores distintos

3.3.1 Publicação do Edital - Análise com Todas as Modalidades

Análise de Variância e de Correlação

Foi feita análise de variância dos dados e análise do Coeficiente de Variação de cada atributo com relação à variável de interesse. Foram eliminados seis atributos.

A verificação da correlação com Coeficiente de Pearson foi feita com o pacote Pandas. Variáveis com correlação superior a 0,7 foram analisadas par a par e atributos com correlações perfeitas foram selecionados e um deles removido com auxílio de especialistas. Após essa etapa três atributos foram eliminados e o conjunto de dados ficou com 369 instâncias e 42 atributos.

Geração de Dados de Avaliação e Balanceamento da Base

Foi gerada uma amostra balanceada e aleatória com total de 34 instâncias e 42 atributos para a etapa de avaliação (Conjunto de dados para avaliação (DAV)), considerando 10% dos dados de cada modalidade. Os dados de avaliação foram removidos do conjunto de treino e teste (Conjunto de para treino/teste dos modelos (DTT)), que ficou com 335 instâncias e 42 atributos, sendo um a variável de interesse.

O conjunto de DTT foi balanceado com uso de SMOTE e passou a contar com 346 instâncias.

Seleção de Variáveis

Foram aplicados os métodos de seleção de variáveis apresentados na Seção 2.4.4 e o resultado está apresentado na Tabela 3.6, na modalidade "Todas". A seleção com RFE, com algoritmo de Regressão Logística, manteve as 23 variáveis mais importantes. Os demais métodos reduziram de forma considerável o número de atributos. A Figura 3.9 apresenta a distribuição de coeficientes para algumas das variáveis selecionadas.

Tabela 3.6: Resumo aplicação de seleção de variáveis na etapa de Edital.

Modalidade	**Pós. Var. e Corr**	**RFE**	**ElasticNetCV**	**LASSOCV**	**LarsCV**
Todas	42	23	9	8	7
Pregão Eletro./Presen.	40	13	37	14	8
Concorrência/Tomada Pre.	41	20	2	2	3
Dispensa/Inexigibilidade	38	28	4	3	3

3.3.2 Publicação do Edital - Pregão Eletrônico e Pregão Presencial

Conforme demonstrado na Figura 3.4, o conjunto de dados corresponde a todas licitações da modalidade pregão eletrônico e pregão presencial, com o total de 172 instâncias e 51 atributos.

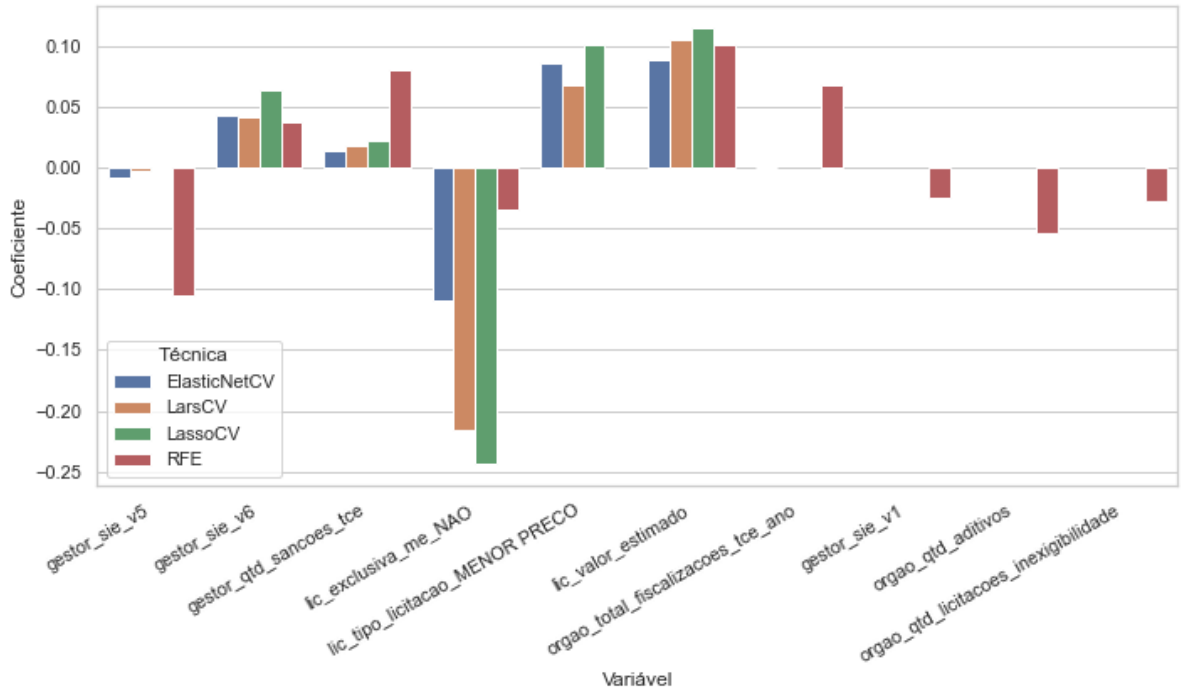

Figura 3.9: Distribuição dos coeficientes por variável selecionada.

Fonte: Próprio autor

Análise de Variância e de Correlação

Foi feita análise de variância dos dados e análise do Coeficiente de Variação de cada atributo com relação à variável de interesse. Foram eliminados oito atributos.

A verificação da correlação com Coeficiente de Pearson foi feita com o pacote Pandas. Variáveis com correlação superior a 0,7 foram analisadas par a par e atributos com correlações perfeitas foram selecionados e um deles removido com auxílio de especialistas. Após essa etapa, três atributos foram eliminados e o conjunto de dados ficou com 172 instâncias e 40 atributos.

Geração de Dados de Avaliação e Balanceamento da Base

Foi gerada uma amostra balanceada e aleatória com total de 15 instâncias e 40 atributos para a etapa de avaliação (DAV). Essas foram removidas do conjunto de treino e teste (DTT), que ficou com 157 instâncias e 40 atributos.

O conjunto de DTT foi balanceado com uso de SMOTE e passou a contar com 166 instâncias.

Seleção de Variáveis

Foram aplicados os métodos de seleção de variáveis apresentados na Seção 2.4.4 e o resultado está apresentado na Tabela 3.6. A seleção com RFE, com algoritmo de Regressão Logística, manteve as 13 variáveis mais importantes, o método *Elastic-Net* com *Cross-Validation* (ElasticNetCV) manteve 37, o LASSOCV manteve 15 e o *Least Angle Regression* com *Cross-Validation* (LarsCV) apenas 8. A Figura 3.10 apresenta a distribuição de coeficientes para algumas das variáveis selecionadas.

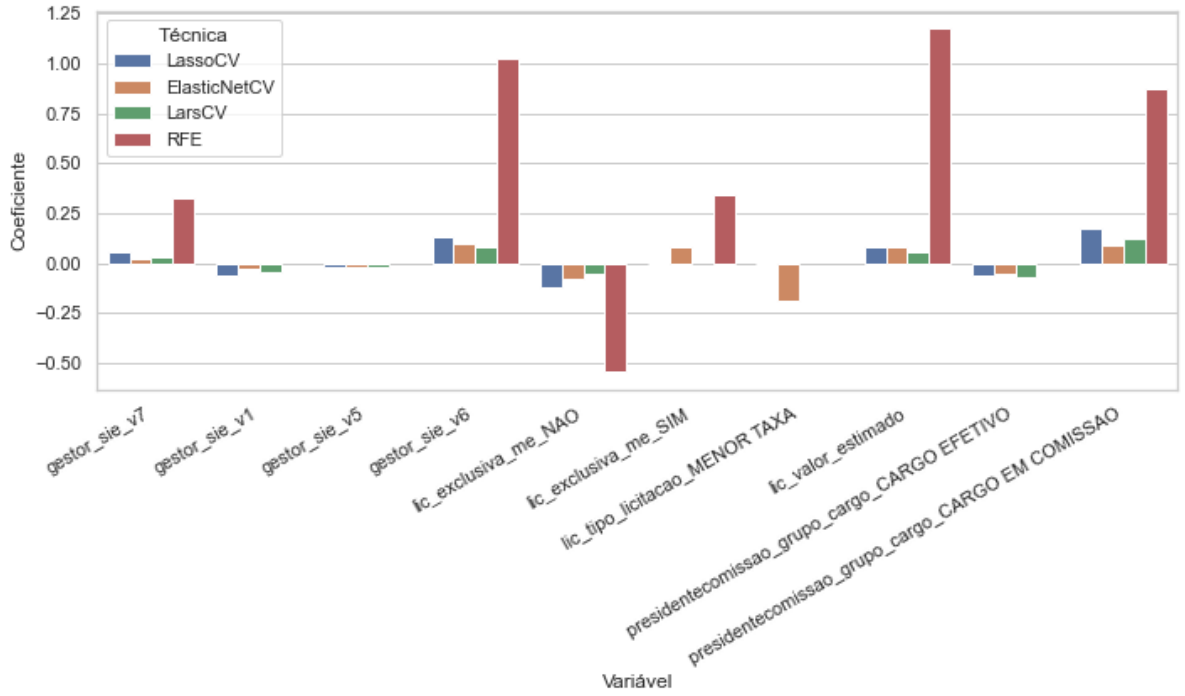

Figura 3.10: Distribuição dos coeficientes por variável selecionada - Pregão.

Fonte: Próprio autor

3.3.3 Publicação do Edital - Concorrência e Tomada de Preço

Conforme demonstrado na Figura 3.4, o conjunto de dados para esta análise corresponde a todas licitações da modalidade concorrência e tomada de preço, totalizando 145 instâncias e 51 atributos.

Análise de Variância e Correlação

Foi feita análise de variância dos dados e análise do Coeficiente de Variação de cada atributo com relação à variável de interesse. Foram eliminados oito atributos.

A verificação da correlação com Coeficiente de Pearson foi feita com o pacote Pandas. Variáveis com correlação superior a 0,7 foram analisadas par a par e atributos com correlações perfeitas foram selecionados e um deles removido com auxílio de especialistas. Após essa etapa, 2 atributos foram eliminados e o conjunto de dados ficou com 145 instâncias e 41 atributos.

Geração de Dados de Avaliação e Balanceamento da Base

Foi gerada uma amostra balanceada e aleatória com total de 15 instâncias e 41 atributos para a etapa de avaliação (DAV). Esses registros foram removidos do conjunto de treino e teste (DTT), que ficou com 130 instâncias e 41 atributos.

O conjunto DTT foi balanceado com uso de SMOTE e passou a contar com 150 instâncias.

Seleção de Variáveis

Foram aplicados os métodos de seleção de variáveis apresentados na Seção 2.4.4 e o resultado está apresentado na Tabela 3.6. A seleção com RFE, com algoritmo de Regressão Logística, manteve as 20 variáveis mais importantes, mas os outros métodos removeram praticamente todas as variáveis. A Figura 3.11 apresenta a distribuição de coeficientes para as variáveis.

3.3.4 Publicação do Edital - Dispensa de Licitação e Inexigibilidade

Conforme demonstrado na Figura 3.4, o conjunto de dados para esta análise corresponde a todas licitações de dispensa e inexigibilidade, totalizando 52 instâncias e 51 atributos.

Análise de Variância e Correlação

Foi feita análise de variância dos dados e análise do Coeficiente de Variação de cada atributo com relação à variável de interesse e foram eliminados oito atributos.

A verificação da correlação com Coeficiente de Pearson foi feita com o pacote Pandas. Variáveis com correlação superior a 0,7 foram analisadas par a par e atributos com correlações perfeitas foram selecionados e um deles removido com auxílio de especialistas. Após essa etapa, cinco atributos foram eliminados e o conjunto de dados ficou com 52 instâncias e 38 atributos.

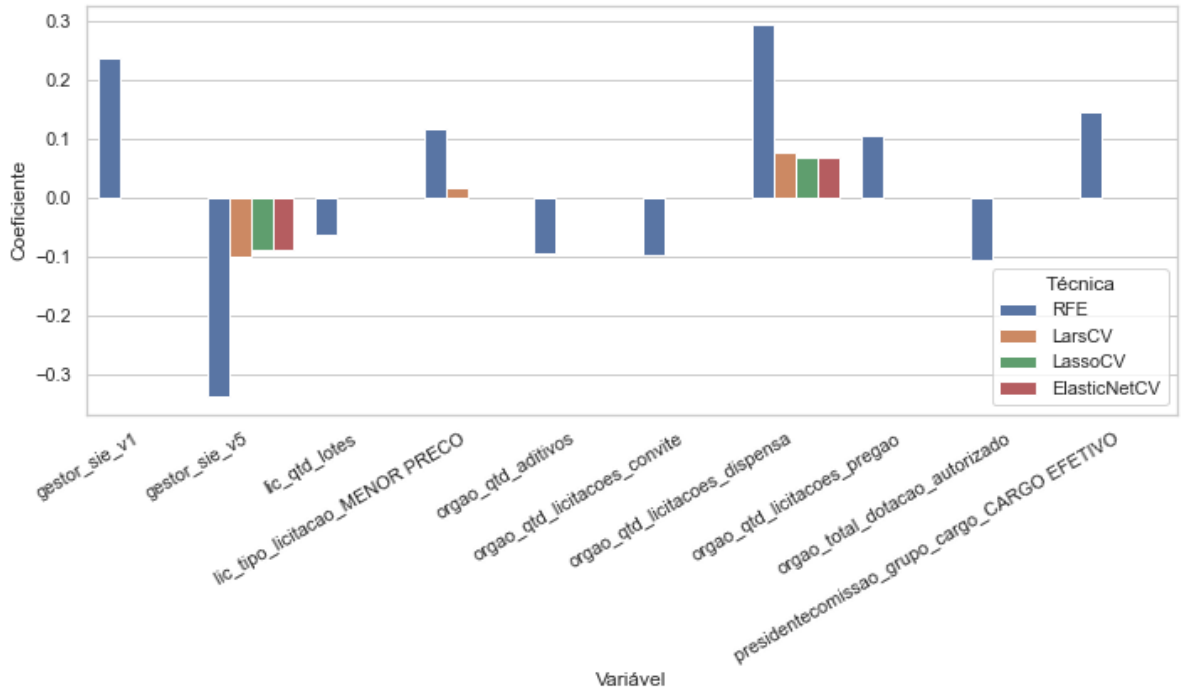

Figura 3.11: Distribuição dos coeficientes por variável selecionada - Concorrência/Tomada de Preço.

Fonte: Próprio autor

Geração de Dados de Avaliação e Balanceamento da Base

Devido à reduzida quantidade de instâncias, não foi gerada amostra de avaliação (DAV). Assim, o conjunto de treino e teste (DTT) ficou com 52 instâncias e 38 atributos, sendo um a variável de interesse.

O conjunto de DTT foi balanceado com uso de SMOTE e passou a contar com 68 instâncias.

Seleção de Variáveis

Foram aplicados os métodos de seleção de variáveis apresentados na Seção 2.4.4 e o resultado está apresentado na Tabela 3.6. A seleção com RFE, com algoritmo de Regressão Logística, manteve as 28 variáveis mais importantes, mas os outros métodos removeram praticamente todas as variáveis. A Figura 3.12 apresenta a distribuição de coeficientes para as variáveis.

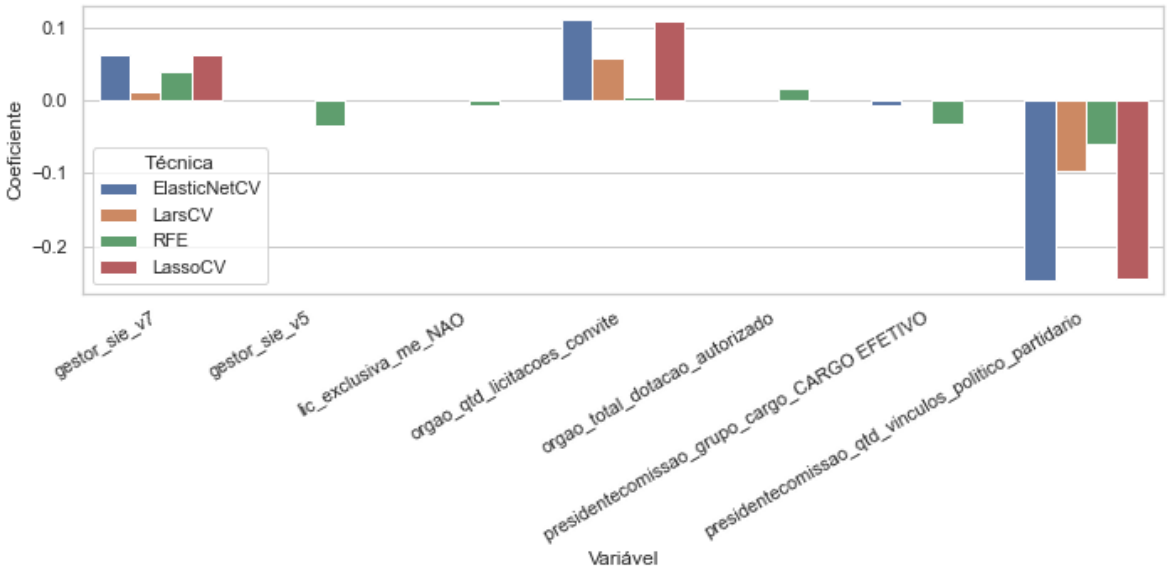

Figura 3.12: Distribuição dos coeficientes por variável selecionada - Dispensa e Inexigibilidade.

Fonte: Próprio autor

3.3.5 Fase de Disputa - Análise com todas as Modalidades

O conjunto de dados para essa etapa é composto por 548 instâncias, com 102 atributos.

Tabela 3.7: Resumo aplicação de seleção de variáveis na etapa de Disputa.

Modalidade	Pós. Var. e Corr	RFE	ElasticNetCV	LASSOCV	LarsCV
Todas	87	41	28	28	10
Pregão Eletro./Presen.	75	43	79	34	15
Concorrência/Tomada Pre.	83	39	7	7	7
Dispensa/Inexigibilidade	60	30	38	9	5

Análise de Variância e Correlação

Foi feita análise de variância dos dados e análise do Coeficiente de Variação de cada atributo com relação à variável de interesse. Foram eliminados sete atributos.

A verificação da correlação com Coeficiente de Pearson foi feita com o pacote Pandas. Variáveis com correlação superior a 0,7 foram analisadas par a par e atributos com correlações perfeitas foram selecionados e um deles removido com auxílio de especialistas. Após essa etapa, oito atributos foram eliminados e o conjunto de dados ficou com 548 instâncias e 87 atributos.

Geração de Dados de Avaliação e Balanceamento da Base

Foi gerada uma amostra balanceada e aleatória com 48 instâncias e 87 atributos para a etapa de avaliação (DAV). Os dados de avaliação foram removidos do conjunto de treino e teste (DTT), que ficou com 500 instâncias e 87 atributos, sendo um a variável de interesse.

O conjunto de DTT foi balanceado com uso de SMOTE e passou a contar com 558 instâncias.

Seleção de Variáveis

Foram aplicados os métodos de seleção de variáveis apresentados na Seção 2.4.4 e o resultado está apresentado na Tabela 3.7. A seleção com RFE, com algoritmo de Regressão Logística, manteve 41 variáveis mais importantes, o ElasticNetCV e o LASSOCV mantiveram as mesmas 28 variáveis e o LarsCV restringiu a apenas 10. A Figura 3.13 apresenta a distribuição de coeficientes para parte das variáveis. Nesse gráfico, o ElasticNetCV e o LASSOCV atribuíram pesos semelhantes para as mesmas variáveis.

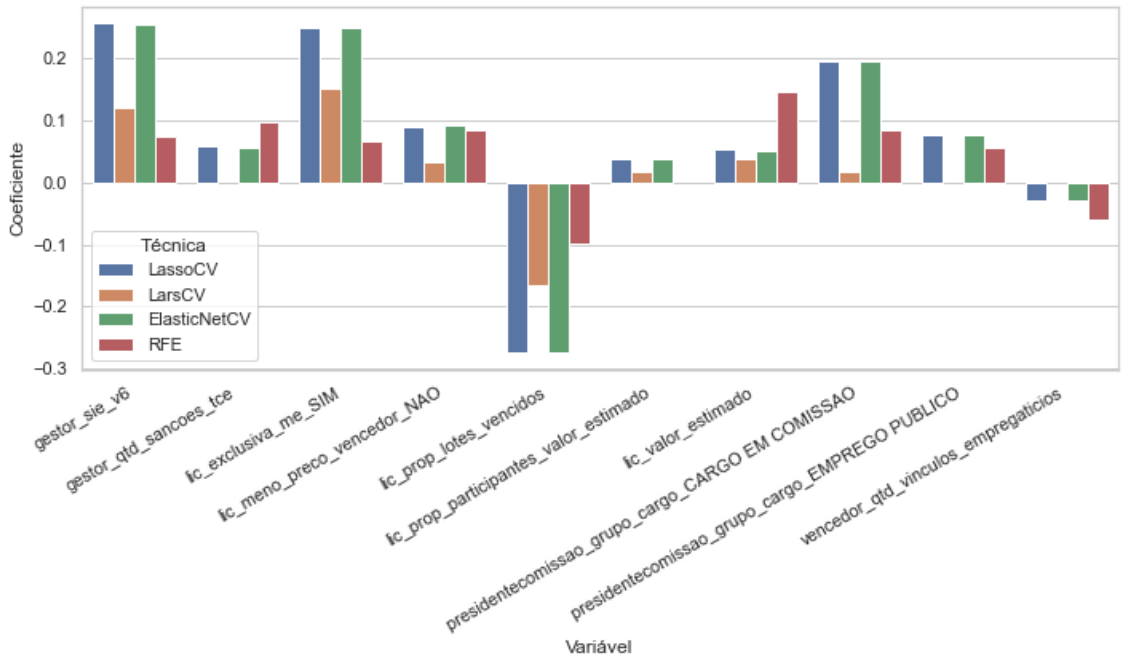

Figura 3.13: Distribuição dos coeficientes por variável selecionada - Todas Modalidades.

Fonte: Próprio autor

3.3.6 Fase de Disputa - Pregão Eletrônico e Pregão Presencial

O conjunto de dados para essa etapa é composto por 345 instâncias, com 102 atributos.

Análise de Variância e Correlação

Foi feita análise de variância dos dados e análise do Coeficiente de Variação de cada atributo com relação à variável de interesse. Foram eliminados nove atributos.

A verificação da correlação com Coeficiente de Pearson foi feita com o pacote Pandas. Variáveis com correlação superior a 0,7 foram analisadas par a par e atributos com correlações perfeitas foram selecionados e um deles removido com auxílio de especialistas. Após essa etapa, 10 atributos foram eliminados e o conjunto de dados ficou com 345 instâncias e 83 atributos.

Geração de Dados de Avaliação e Balanceamento da Base

Foi gerada uma amostra balanceada e aleatória com total de 28 instâncias e 83 atributos para a etapa de avaliação (DAV). Os dados de avaliação foram removidos do conjunto de treino e teste (DTT), que ficou com 317 instâncias e 83 atributos, sendo um a variável de interesse.

O conjunto de DTT foi balanceado com uso de SMOTE e passou a contar com 352 instâncias.

Seleção de Variáveis

Foram aplicados os métodos de seleção de variáveis apresentados na Seção 2.4.4 e o resultado está apresentado na Tabela 3.6. A seleção com RFE, com algoritmo de Regressão Logística, manteve as 43 variáveis mais importantes, o ElasticNetCV selecionou 79, o LASSOCV manteve 34 variáveis e o LarsCV restringiu a apenas 15. A Figura 3.14 apresenta a distribuição de coeficientes para parte das variáveis. Nesse gráfico, o ElasticNetCV e o LASSOCV atribuíram pesos semelhantes às mesmas variáveis enquanto que para o RFE os atributos “lic_valor_estimado”, ”lic_meno_preco_vencedor” e ”gestor_sie_v7” apresentou o maior coeficiente positivo em relação aos outros métodos.

3.3.7 Fase de Disputa - Concorrência e Tomada de Preço

O conjunto de dados para essa etapa é composto por 149 instâncias, com 102 atributos.

Análise de Variância e Correlação

Foi feita análise de variância dos dados e análise do Coeficiente de Variação de cada atributo com relação à variável de interesse. Foram eliminados 10 atributos.

A verificação da correlação com Coeficiente de Pearson foi feita com o pacote Pandas. Variáveis com correlação superior a 0,7 foram analisadas par a par e atributos com correla-

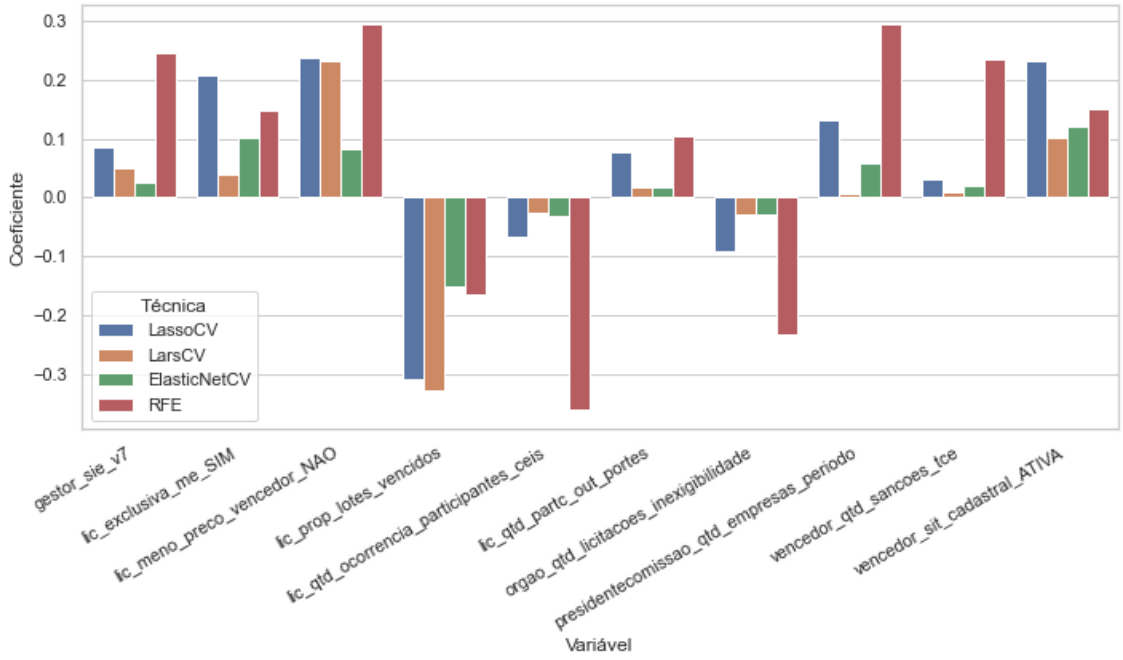

Figura 3.14: Distribuição dos coeficientes por variável selecionada - Pregão.

Fonte: Próprio autor

ções perfeitas foram selecionados e um deles removido com auxílio de especialistas. Após essa etapa, 17 atributos foram eliminados e o conjunto de dados ficou com 149 instâncias e 75 atributos.

Geração de Dados de Avaliação e Balanceamento da Base

Foi gerada uma amostra balanceada e aleatória com total de 10 instâncias e 75 atributos para a etapa de avaliação (DAV). Os dados de avaliação foram removidos do conjunto de treino e teste (DTT), que ficou com 139 instâncias e 75 atributos, sendo um a variável de interesse.

O conjunto de DTT foi balanceado com uso de SMOTE e passou a contar com 174 instâncias.

Seleção de Variáveis

Foram aplicados os métodos de seleção de variáveis apresentados na Seção 2.4.4 e o resultado está apresentado na Tabela 3.7. A seleção com RFE, com algoritmo de Regressão Logística, manteve 39 variáveis mais importantes e todas as demais mantiveram sete atributos. A Figura 3.15 apresenta a distribuição de coeficientes para parte das variáveis.

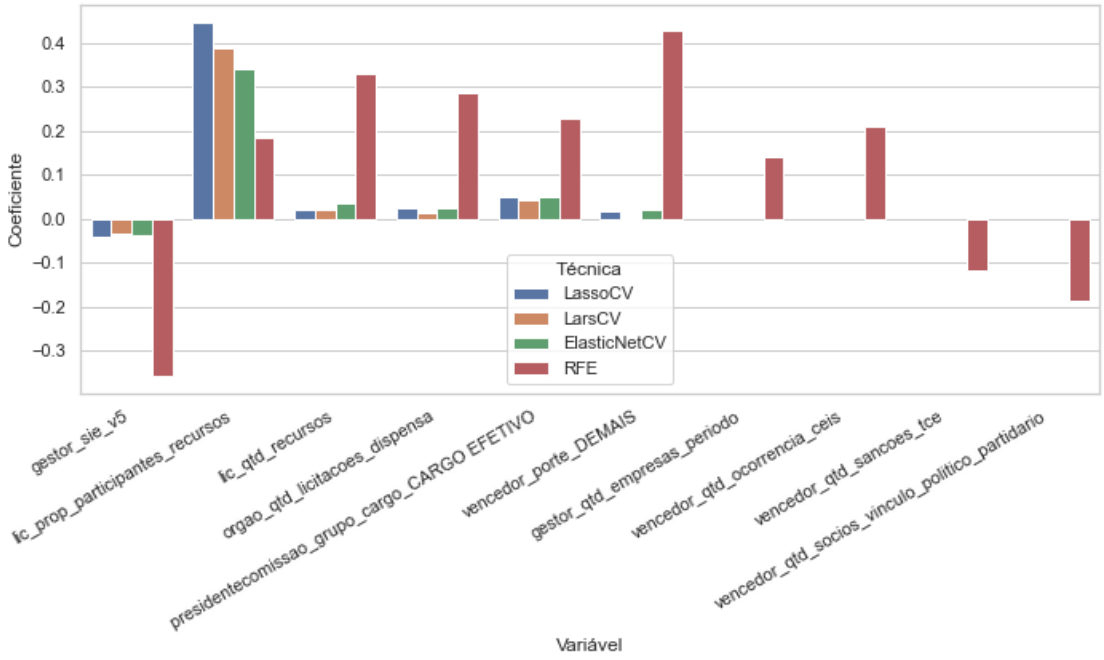

Figura 3.15: Distribuição dos coeficientes por variável selecionada - concorrência e tomada de preço.

Fonte: Próprio autor

3.3.8 Fase de Disputa - Dispensa de Licitação e Inexigibilidade

O conjunto de dados para essa etapa é composto por 54 instâncias, com 102 atributos.

Análise de Variância e Correlação

Foi feita análise de variância dos dados e análise do Coeficiente de Variação de cada atributo com relação à variável de interesse. Foram eliminados 18 atributos.

A verificação da correlação com Coeficiente de Pearson foi feita com o pacote Pandas. Variáveis com correlação superior a 0,7 foram analisadas par a par e atributos com correlações perfeitas foram selecionados e um deles removido com auxílio de especialistas. Após essa etapa, 24 atributos foram eliminados e o conjunto de dados ficou com 54 instâncias e 60 atributos.

Geração de Dados de Avaliação e Balanceamento da Base

Devido à reduzida quantidade de instâncias, não foi gerada amostra de avaliação (DAV). Assim, o conjunto de treino e teste (DTT) ficou com 54 instâncias e 60 atributos, sendo um a variável de interesse.

O conjunto de DTT foi balanceado com uso de SMOTE e passou a contar com 72 instâncias.

Seleção de Variáveis

Foram aplicados os métodos de seleção de variáveis apresentados na Seção 2.4.4 e o resultado está apresentado na Tabela 3.6. A seleção com RFE, com algoritmo de Regressão Logística, manteve 37 variáveis mais importantes e todas as demais mantiveram sete atributos. A Figura 3.15 apresenta a distribuição de coeficientes para parte das variáveis. Nesse gráfico, ElasticNetCV, LASSOCV e LarsCV atribuíram pesos semelhantes às mesmas variáveis enquanto que para o RFE o atributo "lic_qtd_recursos"apresentou o maior valor absoluto de coeficiente em relação aos outros métodos. Além disso, a maioria das variáveis apresentaram coeficientes com valor absoluto pequeno.

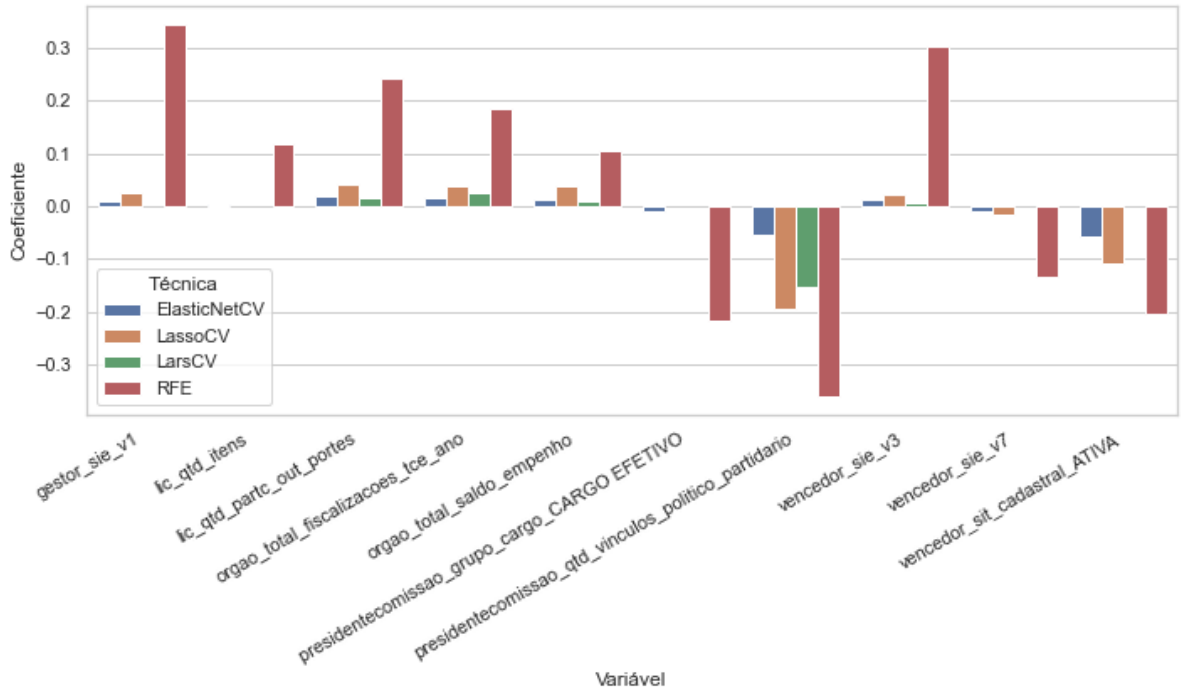

Figura 3.16: Distribuição dos coeficientes por variável selecionada - dispensa e inexigibilidade.

Fonte: Próprio autor

3.4 Modelagem

Esta etapa utiliza como insumo os dados processados nas etapas anteriores e atributos selecionados para treinamento supervisionado de modelos de classificação. O processo de modelagem é feito oito vezes, sendo organizado por fase da licitação (Edital/Disputa) e por conjunto de dados de acordo com a modalidade (Todas Modalidades; Pregão; Concorrência/Tomada de Preços; Dispensa/Inexigibilidade), conforme apresentado na Figura 3.3.

1. Para cada Fase, criou-se quatro conjunto de dados (DCM), sendo (DCMa) um conjunto com todos os dados disponíveis e outros três com dados agrupados por modalidades, sendo (DCMb) pregão presencial e pregão eletrônico, (DCMc) concorrência e tomada de preço e (DCMd) dispensa e inexigibilidade.

2. Para cada conjunto de dados (DCM), aplicaram-se as técnicas de seleção de variáveis e cálculo dos coeficientes apresentados na Seção 3.3, gerando subconjuntos dados com atributos selecionados (DSV) para cada técnica de seleção aplicada.

3. Cada conjunto DSV foi divido aleatoriamente em conjunto de treino/teste (DTT) e conjunto de avaliação (DDV), sendo DDV balanceado e de tamanho aproximadamente de 10% de DSV. Essa etapa não foi realizada para Dispensa devido ao reduzido tamanho das amostras.

4. O conjunto de DTT foi utilizado para treinamento e teste dos modelos de classificação, utilizando *cross-validation* com k igual a 10. O classificador foi treinado e testado uma vez com DDT e outra vez com DDT aplicando-se PCA, com o objetivo de verificar se o uso de componentes principais melhoraria o poder de predição dos classificadores.

5. Cada classificador foi ajustado com hiper parâmetros utilizando técnicas de *Pipeline* e *Grid-Search*. No *Pipeline* foram definidas as etapas de processamento a que cada classificador foi submetido e no *Grid-Search* foram inseridas as combinações possíveis de parâmetros de ajustes dos modelos.

6. O processo de escolha da melhor combinação de parâmetros foi feito com uso de *cross-validation* com k igual a 10 e a métrica de decisão foi a acurácia. Também foram calculadas as métricas de *Precision*, *Recall*, *F1* e *AUROC*.

7. Os melhores modelos treinados foram ranqueados pela acurácia e selecionados. Utilizando os parâmetros ajustados, foi feito o treinamento desses modelos com conjunto DTT e o modelo treinado foi utilizado para uma predição com os dados DDV.

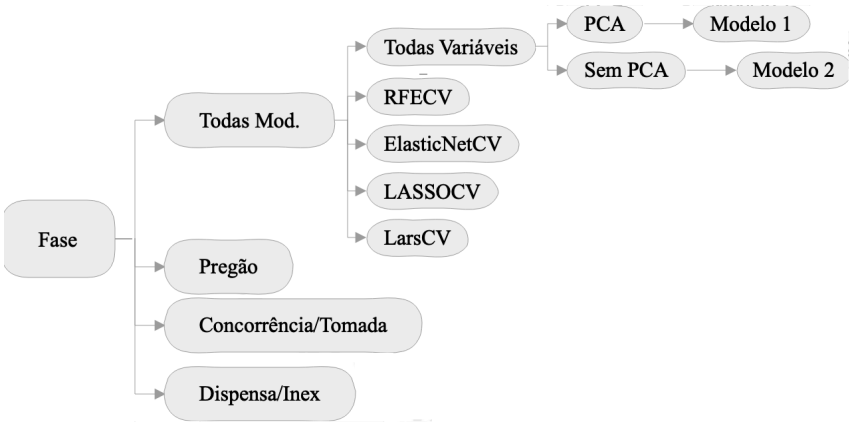

Figura 3.17: Processo de modelagem por fase, modalidade e algoritmo de seleção de variáveis.

Fonte: Próprio autor

A Figura 3.17 apresenta uma visão esquemática do processo de modelagem realizado.

Os classificadores testados possuem diferentes parâmetros que podem ser ajustados para melhorar o poder de predição. Assim, foram realizadas diferentes combinações de parâmetros com *GridSearch* da biblioteca *Scikit-Learn*, conforme apresentado na Tabela 3.8. Os resultados da etapa de modelagem são discutidos na Seção 4.1.

Tabela 3.8: Configuração de parâmetros para os classificadores.

Classificador	**Parâmetro**	**Valores**
RandomForestClassifier	n_estimators	[100, 300, 500, 800]
	max_depth	[5, 8, 15, 25, 30, 50]
	min_samples_split	[2, 5, 10, 15]
	min_samples_leaf	[1, 2, 5, 10]
SVM	kernel	[rbf]
	C	[0.01 , 0.1 , 1, 1.1, 1.274, 2.274]
LogisticRegression	penalty	[l1,l2,elasticnet,none]
	C	[0.01 , 0.1 , 1, 1.1, 1.274, 2.274]
	solver	[liblinear,lbfgs]
XGBClassifier	learning_rate	[0.1, 0.2,0.02]
	min_child_weight	[1, 5, 10]
	gamma	[0.5, 1, 2, 5]
	subsample	[0.6, 0.8, 1.0]
	colsample_bytree	[0.6, 0.8, 1.0]
	max_depth	[3, 4, 5]

Capítulo 4

Resultados

4.1 Avaliação dos Resultados

Neste capítulo são apresentados os resultados obtidos e apresentadas as métricas para verificar a qualidade dos modelos.

4.1.1 Fase de Publicação do Edital

Modelo Geral - Utilizando todo o conjunto de dados

Os modelos treinados na Seção 3.4 para a fase de publicação de edital foram ordenados pela resposta da acurácia, obtida via *cross-validation* com k igual 10. A Tabela 4.1 apresenta os melhores modelos considerando todo o conjunto de dados da desta etapa, além de outras medidas de validação para melhor análise.

Tabela 4.1: Melhores resultados com todo conjunto de dados na fase de edital

Modelo	**Nr. Atributos**	**Acurácia**	**Precision**	**F1**	**Recall**	**AUROC**
RFC_ElasticNetCV	9	61,92%	64,87%	61,10%	62,45%	61,94%
XGB_LassoCV	8	60,74%	61,42%	60,45%	62,06%	59,95%
PCA_RFC_LassoCV	8	60,48%	65,62%	59,10%	59,60%	62,37%

Conforme apresentado na Tabela 4.1, o melhor resultado de acurácia foi de apenas 61,92% para *Random Forest*, utilizando os atributos selecionados por ElasticNetCV. Além disso, todos os classificadores tiveram AUROC abaixo de 65%, o que desencoraja seu uso em ambiente de produção na forma que se encontra. Utilizando o conjunto de dados de avaliação (DDA), foi verificada a assertividade dos modelos, conforme a modalidade e o resultado apresentado na Tabela 4.2.

A Tabela 4.2 corrobora com os indicadores apresentados na Tabela 4.1. Utilizando-se todos os dados, nenhum classificador alcançou mais do que 60% de asserto. Analisando por

Tabela 4.2: Percentual de assertividade com dados de avaliação.

Modelo	% Todas Mod.	% Pregao	% Concorrencia	% Dispensa
RFC_ElasticNetCV	52,94%	62,50%	30,00%	75,00%
XGB_LassoCV	55,88%	56,25%	40,00%	75,00%
PCA_RFC_LassoCV	58,82%	68,75%	30,00%	75,00%

modalidade, as modalidades dispensa e inexigibilidade possuem os melhores resultados, e concorrência e tomada de preço os piores com apenas 30%. Diante desse resultado, conclui-se que um modelo geral para a fase de edital não é adequado.

Modelo Edital-Pregão

A Tabela 4.3 apresenta os três melhores classificadores e técnica de seleção de atributos, ordenados pela métrica de acurácia. Percebe-se uma ligeira melhora com relação ao modelo geral com todos os dados e na assertividade com dados de avaliação. O melhor classificador foi *Radom Forest* com LASSOCV com 14 atributos, pois apresentou os melhores resultados para a maioria das métricas. Os parâmetros ajustados para *Radom Forest* foram os seguintes:

- max_depth': 30
- min_samples_leaf: 1
- min_samples_split: 5
- n_estimators: 300

Tabela 4.3: Melhores modelos para etapa de edital com modalidade pregão

Modelo	Nr. Atributos	Acurácia	Precision	F1	Recall	AUROC	% Avaliação
RFC_LassoCV	14	64,93%	74,01%	65,37%	67,50%	71,13%	73,33%
RFC_TODAS_VARIAVEIS	37	64,19%	75,82%	62,55%	61,94%	70,25%	66,67%
RFC_ElasticNetCV	37	63,05%	74,61%	62,01%	61,11%	68,28%	73,33%

Foi verificado se as métricas de avaliação possuem distribuição normal no conjunto de dados, através de *Bootstrap* com mil amostras aleatórias do mesmo tamanho do conjunto de dados e com repetição. Os histogramas apresentados na Figura 4.1 demostram que todas as métricas possuem distribuição aproximadamente normal, um indicativo de que os resultados apresentados são adequados.

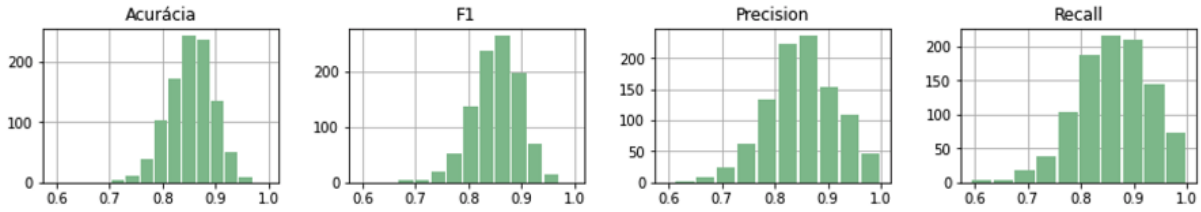

Figura 4.1: Histogramas com distribuição Bootstrap para fase Edital-Pregão

Fonte: Próprio autor

Para esse modelo, a aplicação de LASSOCV selecionou 14 atributos, conforme a Tabela 4.4.

Tabela 4.4: Atributos selecionados para Edital-Pregão

Atributo	**Coeficiente**
gestor_sie_v3	0,007569
gestor_sie_v7	0,055298
gestor_sie_v1	-0,060002
gestor_sie_v5	-0,017911
gestor_sie_v6	0,127645
lic_exclusiva_me_NAO	-0,120487
lic_exclusiva_me_SIM	0,059318
lic_tipo_licitacao_MAIOR OFERTA	0,022957
lic_tipo_licitacao_MENOR TAXA	-0,004283
lic_valor_estimado	0,082366
orgao_qtd_licitacoes_convite	-0,011244
orgao_total_dotacao_autorizado	-0,005856
presidentecomissao_grupo_cargo_CARGO EFETIVO	-0,063208
presidentecomissao_grupo_cargo_CARGO EM COMISSAO	0,169096

As variáveis selecionadas são de quatro bases de dados, todas de acesso permanente pelo TCE.

Modelo Edital-Concorrência

A Tabela 4.5 apresenta os melhores resultados alcançados considerando apenas a modelagem com modalidade concorrência e tomada de preço.

Tabela 4.5: Melhores modelos para etapa de edital com modalidade concorrência ou tomada de preço

Modelo	Nr. Atributos	Acurácia	Precision	F1	Recall	AUROC	% Avaliação
XGB_TODAS_VARIAVEIS	38	65,33%	67,80%	63,25%	62,14%	64,29%	45,00%
XGB_RFE	20	65,33%	67,04%	58,83%	57,86%	61,79%	60,00%
RFC_RFE	20	60,66%	68,83%	56,90%	57,85%	66,16%	55,00%

Assim como aconteceu com a modelagem com todos os dados, os resultados para concorrência ficaram inferiores a 70% em todas as métricas avaliadas. O desempenho com dados de avaliação também foi baixo, o que desencoraja o uso de qualquer um desses modelos na forma como se encontram. Assim, conclui-se que não foi possível obter um modelo confiável na fase de edital para modalidade concorrência ou tomada de preços.

Modelo Edital-Dispensa

A Tabela 4.6 apresenta os melhores resultados alcançados considerando apenas a modelagem com casos de dispensa e inexigibilidade de licitação. Por questões de simplicidade, o classificador *Support Vector Machine* com 28 variáveis selecionadas por RFE, apresentou os melhores resultados para a maioria das métricas observadas. Os parâmetros ajustados para o *Support Vector Machine* foram os seguintes:

- C: 1.274
- kernel: rbf

Tabela 4.6: Melhores modelos para etapa de edital com dispensa e inexigibilidade

Modelo	Nr. Atributos	Acurácia	Precision	F1	Recall	AUROC
PCA_SVM_RFE	28	72,14%	88,09%	68,95%	73,33%	81,39%
PCA_SVM_TODAS_VARIAVEIS	34	72,14%	85,72%	68,12%	70,83%	80,00%
SVM_RFE	28	70,71%	87,50%	67,62%	74,17%	82,22%

Foi verificado se as métricas de avaliação possuem distribuição normal no conjunto de dados, através de *Bootstrap* com mil amostras aleatórias do mesmo tamanho do conjunto de dados e com repetição. Os histogramas apresentados na Figura 4.2 demostram que todas as métricas possuem distribuição aproximadamente normal, um indicativo de que os resultados apresentados são adequados.

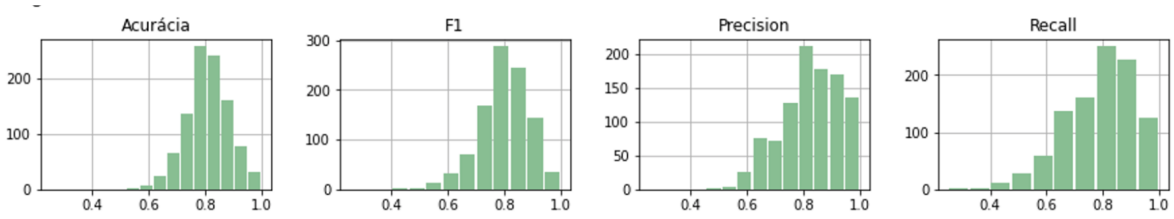

Figura 4.2: Histogramas com distribuição *Bootstrap* para fase Edital-Dispensa

Fonte: Próprio autor

Para esse modelo, a aplicação de RFE selecionou 28 atributos, conforme a Tabela 4.7.

Tabela 4.7: Atributos selecionados para Edital-Dispensa.

Atributo	**Coeficiente**
orgao_total_fiscalizacoes_tce_ano	0,066861
gestor_sie_v1	0,066792
presidentecomissao_qtd_vinculos_politico_partidario	-0,060197
gestor_qtd_sancoes_tce	0,055844
gestor_qtd_empresas_periodo	0,051026
lic_qtd_lotes	-0,047308
orgao_qtd_licitacoes_inexigibilidade	0,047049
lic_valor_estimado	0,04402
orgao_total_saldo_empenho	0,042913
gestor_sie_v7	0,040028
gestor_sie_v5	-0,033589
presidentecomissao_qtd_empresas_periodo	-0,033189
presidentecomissao_grupo_cargo_CARGO EFETIVO	-0,03085
presidentecomissao_qtd_sancoes_tce	0,027043
orgao_qtd_licitacoes_dispensa	-0,027038
gestor_qtd_vinculos_politico_partidario	-0,02694
gestor_grupo_cargo_CARGO EM COMISSAO	-0,024846
orgao_qtd_licitacoes_pregao	-0,023627
lic_qtd_itens	0,021058
orgao_qtd_licitacoes_concorrencia	0,0183
orgao_total_dotacao_autorizado	0,016986
gestor_grupo_cargo_EMPREGO PUBLICO	-0,013521

Tabela 4.7 continuação da página anterior.

Atributo	Coeficiente
orgao_qtd_licitacoes	-0,011211
presidentecomissao_grupo_cargo_EMPREGO PUBLICO	-0,009355
orgao_qtd_licitacoes_convite	0,006684
lic_exclusiva_me_NAO	-0,006307
orgao_total_dotacao_cred_especial	0,00572
lic_tipo_licitacao_LOCACAO IMOVEL	-0,004507

4.1.2 Fase de Disputa

Modelo Disputa-Geral - Utilizando todo o conjunto de dados

Os modelos treinados na Seção 3.4 para a fase de disputa foram ordenados pela resposta da acurácia, obtida via *cross-validation* com k igual 10. A Tabela 4.8 apresenta os melhores modelos considerando todo o conjunto de dados desta etapa, e diversas medidas de validação para melhor análise.

Tabela 4.8: Melhores resultados com todo conjunto de dados na fase de disputa

Modelo	Nr. Atributos	Acurácia	Precision	F1	Recall	AUROC
RFC_LassoCV	28	69,02%	78,84%	63,42%	56,31%	75,65%
RFC_RFE	43	68,49%	78,46%	63,29%	56,28%	76,71%
XGB_RFE	43	67,93%	76,25%	64,54%	60,89%	73,92%

Conforme apresentado na Tabela 4.8, os modelos possuem valores próximos para a maioria das métricas. Nenhum modelo conseguiu acurácia, F1 ou Recall acima de 70%. Utilizando o conjunto de dados de avaliação (DDA), foi verificada a assertividade dos modelos conforme a modalidade e o resultado é apresentado na Tabela 4.9.

Tabela 4.9: Percentual de assertividade com dados de avaliação.

Modelo	%Todas Mod.	%Pregão	% Concorrência	% Dispensa
RFC_LassoCV	68,75%	75,86%	60,00%	55,60%
RFC_RFE	66,67%	72,41%	60,00%	66,67%
XGB_RFE	72,92%	75,86%	50,00%	88,89%

A Tabela 4.9 corrobora com os indicadores apresentados na Tabela 4.8. Utilizando-se todos os dados, apenas XGB com RFE obteve resultado acima de 70% de asserto. Em geral, a modalidade pregão possui melhores resultados do que concorrência e dispensa,

mas os resultados demostram que não há um classificador com desempenho acima de 70% para todas as modalidades simultaneamente. Diante desse resultado, conclui-se que um modelo geral para a fase de disputa não é adequado na forma em que se encontra.

Modelo Disputa-Pregão

A Tabela 4.10 apresenta os melhores classificadores e técnicas de seleção de atributos, ordenados pela métrica de acurácia, calculada por meio de *cross-validation* com k igual a 10. Percebe-se uma significativa melhora com relação ao modelo geral com todos os dados e considerável aumento de acurácia com dados de avaliação. Os classificadores apresentam resultados próximos para a maioria das métricas, mas por questão de simplicidade, o modelo *XGB* com LarsCV com 15 atributos foi selecionado. Os parâmetros ajustados para *XGB* são os seguintes:

- max_depth: 30
- min_samples_leaf: 1
- min_samples_split: 5
- n_estimators: 300

Tabela 4.10: Melhores modelos para etapa de disputa com modalidade pregão

Modelo	Nr. Atributos	Acurácia	Precision	F1	Recall	AUROC	% Avaliação
XGB_LarsCV	15	75,84%	81,57%	71,52%	71,05%	76,67%	89,29%
XGB_TODAS_VARIAVEIS	82	74,96%	83,40%	71,01%	68,73%	80,07%	92,86%
RFC_LarsCV	15	74,41%	84,35%	69,90%	68,17%	79,56%	96,43%

Foi verificado se as métricas de avaliação possuem distribuição normal no conjunto de dados, através de *Bootstrap* com mil amostras aleatórias do mesmo tamanho do conjunto de dados e com repetição. Os histogramas apresentados na Figura 4.3 demostram que todas as métricas possuem distribuição aproximadamente normal, um indicativo de que os resultados apresentados são adequados.

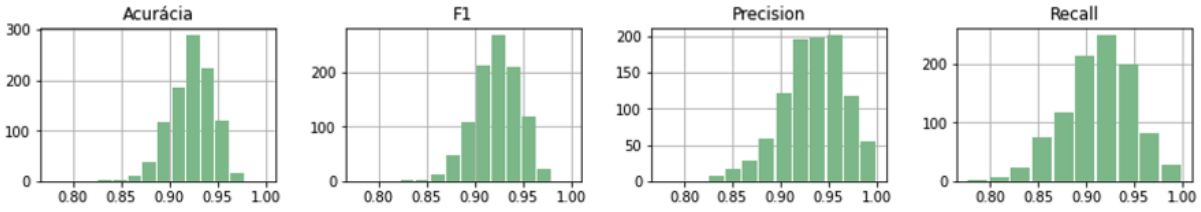

Figura 4.3: Histogramas com distribuição Bootstrap para fase Disputa-Pregão

Fonte: Próprio autor

Para esse modelo, a aplicação de LarsCV selecionou 15 atributos, conforme a Tabela 4.11.

Tabela 4.11: Atributos selecionados para Disputa-Pregão

Atributo	**Coeficiente**
gestor_grupo_cargo_CARGO EFETIVO	-0,001852344
gestor_sie_v7	0,050783289
gestor_qtd_vinculos_politico_partidario	0,014293474
lic_exclusiva_me_SIM	0,03811929
lic_meno_preco_vencedor_NAO	0,230545851
lic_prop_lotes_vencidos	-0,328841238
lic_qtd_ocorrencia_participantes_ceis	-0,02538122
lic_qtd_partc_out_portes	0,016988187
orgao_qtd_licitacoes_inexigibilidade	-0,028519602
presidentecomissao_grupo_cargo_CARGO EFETIVO	-0,058572104
presidentecomissao_qtd_empresas_periodo	0,005822623
presidentecomissao_qtd_sancoes_tce	-0,016931052
vencedor_porte_DEMAIS	0,032241767
vencedor_qtd_sancoes_tce	0,009408986
vencedor_sit_cadastral_ATIVA	0,101840658

Foi constatado uma redução de 81% das variáveis originais. O conjunto de variáveis para pregão na etapa de disputa foi diferente do conjunto variáveis na etapa de edital. Aqui se percebe forte influência de atributos do gestor do órgão e do presidente da licitação. O maior coeficiente em módulo pertence à variável lic_prop_lotes_vencidos, seguido pela variável categórica lic_meno_preco_vencedor preenchida com valor "não".

Modelo Disputa-Concorrência

A Tabela 4.12 apresenta os melhores classificadores e técnicas de seleção de atributos, ordenados pela métrica de acurácia, calculada por meio de *cross-validation* com k igual a 10. Houve melhora com relação aos indicadores do modelo geral apresentados na Tabela 4.8 e Tabela 4.9. O modelo de Regressão Logística, utilizando PCA com 31 componentes aplicados em 39 atributos selecionados através de RFE apresentou melhor desempenho. Os parâmetros ajustados para Regressão Logística são os seguintes:

- C: 0,01
- penalty: none (sem regularização)

- solver: lbfgs

Tabela 4.12: Melhores modelos para etapa de disputa com modalidade concorrência e tomada de preço

Modelo	Nr. Atributos	Acurácia	Precision	F1	Recall	AUROC	% Avaliação
PCA_LR_RFE	39	70,10%	79,58%	63,68%	62,78%	75,68%	71%
PCA_SVM_TODAS_VARIAVEIS	74	67,09%	75,30%	61,20%	55,69%	70,42%	60%
LR_RFE	39	65,36%	78,07%	59,15%	57,22%	73,83%	55%

Foi verificado se as métricas de avaliação possuem distribuição normal no conjunto de dados, através de *Bootstrap* com mil amostras aleatórias do mesmo tamanho do conjunto de dados e com repetição. Os histogramas apresentados na Figura 4.3 demostram que todas as métricas possuem distribuição aproximadamente normal, um indicativo de que os resultados apresentados são adequados.

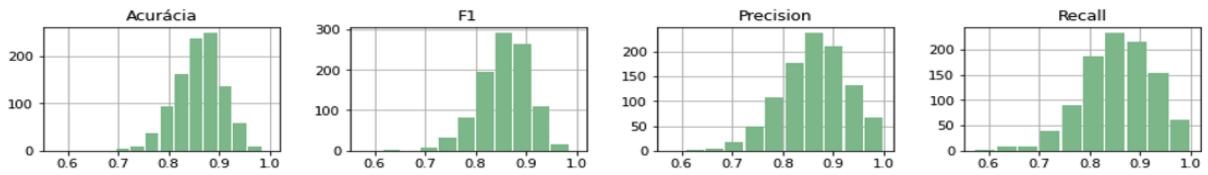

Figura 4.4: Histogramas com distribuição Bootstrap para fase Disputa-Concorrência

Fonte: Próprio autor

Para esse modelo, a aplicação de RFE selecionou 39 atributos, conforme a Tabela 4.13.

Tabela 4.13: Atributos selecionados para Disputa-Concorrência.

Atributo	**Coeficiente**
vencedor_porte_DEMAIS	0,427624
gestor_sie_v5	-0,356174
vencedor_qtd_socios_servidores	0,342712
lic_qtd_recursos	0,330524
orgao_qtd_licitacoes_dispensa	0,288131
orgao_qtd_licitacoes_inexigibilidade	-0,275342
lic_qtd_desclass	0,273783
gestor_sie_v6	0,227968
presidentecomissao_grupo_cargo_CARGO EFETIVO	0,227366

Tabela 4.13 continuação da página anterior.

Atributo	**Coeficiente**
vencedor_sie_v8	0,220653
vencedor_qtd_cnae	-0,219364
vencedor_qtd_ocorrencia_ceis	0,21205
orgao_total_dotacao_cred_especial	0,195176
orgao_total_fiscalizacoes_tce	0,1884
vencedor_qtd_socios_vinculo_politico_partidario	-0,186623
lic_prop_participantes_recursos	0,18603
lic_lotes_vencidos	0,180254
lic_qtd_partc_me	-0,17787
vencedor_qtd_socios	-0,17539
orgao_total_saldo_empenho	-0,168713
lic_qtd_ocorrencia_participantes_ceis	0,167395
orgao_qtd_licitacoes_pregao	0,158237
lic_prop_participantes_desclassificacoes	0,155657
vencedor_sie_v6	-0,153774
vencedor_sie_v1	-0,15195
gestor_qtd_empresas_periodo	0,141122
orgao_qtd_licitacoes	0,138979
vencedor_sie_v3	-0,128754
vencedor_porte_MICRO EMPRESA	0,125634
lic_qtd_ocorrencia_participantes_cnep	0,122625
lic_valor_estimado	0,122124
gestor_sie_v1	0,121031
orgao_qtd_licitacoes_concorrencia	0,118508
vencedor_qtd_sancoes_tce	-0,116498
vencedor_prazo_abertura_ate_licitacao	-0,109057
presidentecomissao_qtd_sancoes_tce	-0,108995
lic_meno_preco_vencedor_NAO	0,10778
participantes_qtd_socios_servidores	0,098098
vencedor_capital_social	0,090357

Modelo Disputa-Dispensa

A Tabela 4.14 apresenta os melhores classificadores e técnicas de seleção de atributos, ordenados pela métrica de acurácia, calculada por meio de *cross-validation* com k igual a

10. Houve melhora com relação aos indicadores do modelo geral apresentados na Tabela 4.8 e Tabela 4.9. O modelo de Regressão Logística aplicado em 30 atributos selecionados através de RFE é utilizado para análise. Os parâmetros ajustados para Regressão Logística são os seguintes:

- C: 1,27427
- penalty: l2
- solver: liblinear

Tabela 4.14: Melhores modelos para etapa de disputa com dispensa e inexigibilidade

Modelo	**Nr. Atributos**	**Acurácia**	**Precision**	**F1**	**Recall**	**AUROC**
LR_RFE	30	84,64%	95,22%	85,44%	87,50%	93,54%
SVM_RFE	30	80,53%	92,95%	81,62%	87,50%	90,83%
SVM_TODAS_VARIAVEIS	59	82,14%	92,82%	81,29%	82,50%	89,58%

Ao se verificar a distribuição das métricas através de *Bootstrap* da Figura 4.5, evidenciou-se que apenas a acurácia segue uma distribuição normal. As métricas *Precision* e *Recall* têm distribuição majoritária acima de 90%, o que pode indicar possível *overfitting*. Foi feita análise das distribuições para os outros três classificadores da Tabela 4.14 e todos têm comportamento similar ao da Regressão Logística. Diante desse resultado, decidiu-se não utilizar esses modelos no estado em que se encontram.

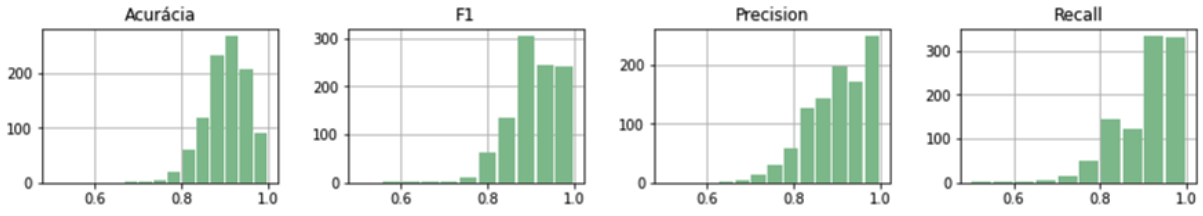

Figura 4.5: Histogramas com distribuição *Bootstrap* para fase Disputa-Dispensa

Fonte: Próprio autor

4.2 Implantação

A proposta da solução foi discutida na Capítulo 3. A Figura 3.1 apresentou a visão conceitual dos modelos treinados aplicados a dados para o estabelecimento de um ranking. Foi construída uma visão na forma de painel dinâmico para validação dos especialistas com licitações na modalidade Pregão, ranqueadas pelo modelo proposto.

Cada licitação tem o risco calculado na etapa de edital (Risco Edital) e na etapa de disputa (Risco Disputa), através do método "predic_proba". Este método faz parte dos estimadores do pacote *Scikit-learn* e retorna a estimativa de probabilidade entre as classes, que varia entre zero (0) e um (1). Assim, observando apenas resultados de classificação para a classe "risco de irregularidade" igual a um (1), retorna-se o valor da estimativa de probabilidade da classificação. Quanto maior o resultado, maior é o risco de irregularidades na licitação. Para facilitar a leitura pelos especialistas, estabeleceu-se três faixas de risco:

1. Risco Alto: valor de risco estimado maior que 71%.
2. Risco Médio: valor de risco estimado entre 51% e 70%.
3. Risco Baixo: valor de risco estimado menor que 51%.

Licitações na fase de disputa podem ter parte dos registros duplicados, dependendo do número de vencedores. Por isso o classificador pode atribuir diferentes percentuais de risco para a mesma licitação. Nessa etapa, considera-se o maior risco encontrado entre os participantes como sendo o risco da licitação.

Risco Licitações

Modalidade | Órgão

Solicitação	Órgão	Objeto	Valor Estimado	Risco Edital	Risco Disputa	Risco Licitação
Totais			**R$223.336.295,60**			
38967	SECRETARIA DE ESTADO DA ADMINISTRACAO	FORNECIMENTO E INSTALAÇÃO ...	R$3.051.843,35	96,42%	82,04%	96,42%
35726	AGENCIA GOIANA DE HABITACAO S/A	SOLICITAÇÃO DE ELABORAÇÃO ...	R$95.672.761,79	91,30%	93,35%	93,35%
35762	AGENCIA GOIANA DE HABITACAO S/A	ELABORAÇÃO DE REGISTRO DE...	R$39.226.954,71	90,19%	93,35%	93,35%
50374	SECRETARIA DE ESTADO DA SAUDE	REGISTRO DE PREÇOS PARA ...	R$10.447.088,00	83,54%	47,77%	83,54%
44052	DEPARTAMENTO ESTADUAL DE TRANSITO	Destina-se a cobrir despesas...	R$12.726.000,00	83,42%	81,42%	83,42%
51473	SECRETARIA DE ESTADO DA ADMINISTRACAO	CONTRATAÇÃO DE EMPRESA ESPECIALIZADA...	R$58.437.500,00	78,74%	82,95%	82,95%
47305	CORPO DE BOMBEIROS MILITAR	CONTRATAÇÃO DE EMPRESA ESPECIALIZADA...	R$1.399.966,90	31,88%	41,77%	41,77%
47076	SECRETARIA DE ESTADO DE AGRICULTURA, PECUARIA E ABASTECIMENTO	CONTRATAÇÃO DE EMPRESA ESPECIALIZADA E...	R$735.211,85	31,88%	15,67%	31,88%
73696	AGENCIA GOIANA DE INFRAESTRUTURA E TRANSPORTES	CONTRATAÇÃO DE EMPRESA PARA ...	R$1.638.969,00	26,94%	22,41%	26,94%

Figura 4.6: Interface em *Qlik Sense* apresentando os riscos estimados para algumas licitações.

Fonte: Próprio autor

O Risco Final é o maior valor valor entre Risco Edital e Risco Disputa. Decidiu-se apresentar as todas medidas de risco para que o auditor do TCE-GO tomasse sua decisão.

A Figura 4.6 apresenta a interface criada na ferramenta *Qlik Sense* do Tribunal. A cor magenta representa licitações com Risco Alto, cor azul com Risco Médio e cor verde Risco Baixo.

Capítulo 5

Conclusão

O presente trabalho apresentou o uso de técnicas de aprendizado supervisionado para se estabelecer um indicador de risco de irregularidades em compras públicas, utilizando dados das compras realizadas no Estado de Goiás.

As decisões do Tribunal de Contas do Estado de Goiás relativas à análise de editais de licitação foram catalogadas e, em seguida, foram identificadas quais compras apresentavam irregularidades e quais não tiveram pendências, segundo análise de auditores do TCE-GO.

O escopo de levantamento de dados é relativo à janeiro de 2014 a dezembro de 2019. Nesse período, foram abertos mais de 1.600 processos de fiscalização de editais de licitação no TCE-GO. Todavia, nesse período, foi possível utilizar 369 licitações para treinamento dos modelos. O desafio de se trabalhar com uma base pequena é real, pois os modelos de *machine learning* necessitam de certa quantidade de instâncias para que consigam realizar aprendizado. Soma-se a isso a quantidade de atributos que existem para predição.

A partir do estudo das normas aplicáveis à licitações e com apoio de especialistas, foram mapeados 87 atributos que potencialmente teriam influência sobre o risco de irregularidades em compras públicas. Foram testadas quatro técnicas de seleção de variáveis e foi constatado que essas técnicas melhoraram o desempenho dos modelos de *machine learning*. Também foram treinados quatro classificadores para comparação de resultados. Todo o processo de mineração de dados foi realizado conforme as fases modelo de referência CRISP-DM.

Inicialmente se buscava um modelo generalista que pudesse estimar o risco nas etapas de publicação edital e na etapa de disputa. Durante as tarefas de modelagem, percebeu-se que o desempenho dos modelos variava de acordo com a modalidade da licitação. Dessa forma, apoiado na pesquisa acadêmica e na legislação relacionada a compras públicas, decidiu-se treinar modelos especialistas por modalidade e comparar o resultado com o modelo geral.

Foi constatado que o modelo geral não foi adequado na maioria dos casos e que modelos especialistas apresentaram melhores resultados de predição. Isso ficou mais evidente nas etapas de seleção de variáveis, pois para cada modalidade, um conjunto distinto de atributos foi selecionado. Esse resultado corrobora com as pesquisas realizadas por Speck & Ferreira [17] e Rodrigues & Notato [16], que apontaram, através de análise jurídica e análise estatística, respectivamente, a relação entre a modalidade de licitação e o risco de ocorrência de fraudes nos processos licitatórios.

Na fase de edital, o modelo especialista para modalidade pregão presencial e pregão eletrônico, utilizando *Random Forest*, apresentou *Area Under the Receiver Operating Characteristic Curve* (AUROC) com 71,13%, com 14 atributos selecionados com Lasso, dos 37 atributos originais. O modelo especialista para dispensa com SVM apresentou AUROC de 81,29%, com 28 atributos selecionados com RFE, dos 34 atributos inciais. Nessa etapa nenhum modelo apresentou bons resultados para concorrência e tomada de preços.

Para a etapa de disputa, o modelo especialista para pregão utilizando *Gradient Boosting* obteve 89,29% de AUROC com 15 tributos selecionados com Lars, dos 82 atributos iniciais. Já o modelo especialista para concorrência e tomada de preço com Regressão Logística apresentou AUROC de 73,83% com 39 atributos selecionados com RFE, dos 74 atributos originais. Para dispensa e inexigibilidade, diversos classificadores apresentaram valores acima de 90% de AUROC. Após análise das distribuições *bootstrap* constatou-se que as medidas de acurácia, F1, *Recall* e precisão não possuem distribuição normal, e, portanto, podem indicar a ocorrência de superajuste.

De modo geral, os modelos que apresentaram bons desempenho foram aqueles que utilizaram redução de dimensionalidade. Nesses casos foram identificados os atributos mais importantes.

Embora os melhores resultados tenham sido alcançados apenas para modalidade pregão eletrônico e pregão presencial nas duas etapas propostas, essas modalidades representaram cerca de R$ 14 bilhões ou 45% do total adjudicado em Goiás nos últimos 5 anos, o que encorajou a proposta de implantação deste modelo para pregão de forma imediata. Para demostrar o funcionamento, foi criado um protótipo de um *ranking* em *Qlik Sense* para que as áreas de negócio pudessem avaliar os resultados. Para cada licitação da modalidade pregão, foi apresentado o Risco Edital e o Risco Disputa.

Outros benefícios dessa pesquisa podem ser enumerados. Com base na análise dos dados dos sistemas, foi constatada a baixa qualidade das informações e discrepâncias que merecem uma olhar atento do controle externo, através de uma proposta de auditoria no sistema Compras-NET do Poder Executivo. Foi constatado que os órgãos não estão atualizando dados no Sistema Informa do TCE-GO, descumprindo resolução do Tribunal. Além disso, com base nos atributos coletados, foi feito ajuste no Sistema SGF do Tribu-

nal para constar dados necessários para o cruzamento de informações com sistemas dos jurisdicionados. Ademais, pretende-se sugerir ajustes no Sistema Informa do Tribunal, incluindo informações que não estão presentes e removendo uma série de outras que não estão agregando valor.

Por fim, foram diversas dificuldades encontradas no decorrer do desenvolvimento deste trabalho. A pandemia da COVID-19 limitou acesso à dados e a especialistas em licitações do Tribunal, bem como dificultou a implantação da solução. Os dados necessários para treinamento são históricos e devem retroagir à data da licitação, mas, em muitos casos, as informações estavam impressas ou digitalizadas de forma não estruturada. Soma-se a isso a pequena quantidade de dados capturados, que tornou a pesquisa mais complexa e forçou busca por métodos de seleção de atributos e ajustes de hiper parâmetros nos classificadores.

Para trabalhos futuros, sugere-se a replicação da pesquisa utilizando uma amostra de dados maior. Modelos especialistas das modalidades concorrência e dispensa tiveram baixos desempenho e isso pode estar relacionado à pequena massa de dados para treinamento. Com uma quantidade maior de dados, pode ser feito estudo da influência de *outliers* nos classificadores. Também podem ser estudadas com maior precisão as relações não lineares entre as variáveis, bem como a influência da colinearidade, caso exista, nas etapas de seleção de atributos baseadas em regressão e também no desempenho dos classificadores. Sugere-se a aplicação de Análise Multicritério, utilizando novos critérios da literatura, pois apenas o risco pode não ser suficiente para tomada de decisão. Nesse caso, outros critérios, tais como valor financeiro, oportunidade de atuação da fiscalização e análise de benefício da ação de controle podem ser considerados.

Referências

[1] Holmes, L.: *Rotten States? Corruption, Post-Communism and Neoliberalism.* Duke University Press, Durham, London, 2006. 1

[2] Remédio, José Antônio; Silva, Marcelo Rodrigues: *Os acordos de leniência da lei anticorrupção e o uso da informação da empresa colaboradora como ativo na reparação integral do dano e no pagamento das sanções pecuniárias.* Revista da AGU, 17:165–184, 2018. 1

[3] Internaconal, Transparência Brasil: *Relatório Executivo 2018.* Relatório Técnico, 2018. `https://transparenciainternacional.org.br/assets/files/conhecimento/relatorio-executivo.pdf`. 1

[4] MPF: *Operação Lava Jato em números*, 2019. 2

[5] AURIOL, EMMANUELLE; STRAUB, STEPHANE; FLOCHEL THOMAS;: *Public Procurement and Rent-Seeking: The Case of Paraguay.* World Development, 77:395–407, 2016. 2

[6] Brasil.: *Constituição da República Federativa do Brasil*, 1988, ISBN 978-85-7018-698-0 1. ISSN 10282092. 2

[7] Batista, Daniel Gerhard: *Manual de Controle e Auditoria.* Editora Saraiva, 2017. 2, 3, 9

[8] GOIÁS: *Constituição do Estado de Goiás*, 1989. 3

[9] Lima, L. H.: *Controle Externo - Teoria e Jurisprudência para os Tribunais de Contas.* Editora Método, São Paulo, º7ºª edição, 2018. 3, 4, 5, 8

[10] Bandiera, Oriana, Andrea Prat e Tommaso Valletti: *Active and passive waste in government spending: Evidence from a policy experiment.* American Economic Review, 2009, ISSN 00028282. 3

[11] Chapman, Pete, Julian Clinton, Randy Kerber, Khabaza Thomas, Thomas Reinartz, Colin Shearer e Wirth Rudiger: *CRISP-DM 1.0. Step-by-step data mining guide.* CRISP-DM Consortium, 2000, ISSN 0957-4174. 3, 6, 18, 19, 28

[12] Balaniuk, R.: *A mineração de dados como apoio ao controle externo.* Revista do TCU, páginas 79–89, 2010. 4

[13] Silva, Marco Aurelio Souza da: *Tribunaide Contas: Teoria e prática da resposabilização de agentes públicos e privados por infração adminitrativa*. Lumen Juris, Rio de Janeiro, 1ª edição, 2017, ISBN 9788551902981. 8

[14] Furtado, Lucas Rocha: *Curso de direito administrativo*. Belo Horizonte, 4ª edição, 2013, ISBN 9788577006786. 10

[15] Carvalho, Matheus: *Manual de direito administrativo*. Editora JusPODIVM, Salvador, 7ª edição, 2019. 10, 11

[16] Rodrigues, N. C. S.; Lima Filho, R. N.: *Modalidades Licitatórias e o Risco De Ocorrência de Fraudes nos Municípios Baianos Fiscalizados Pela Controladoria Geral Da União*. XVI Congresso de Controladoria e Contabilidade da USP, 2017. 10, 12, 13, 30, 82

[17] SPECK, Bruno Wilhelm e Valeriano Mendes FERREIRA: *Sistemas de integridade nos estados brasileiros*. São Paulo: Instituto Ethos de Empresas e Responsabilidade Social, 2012. 12, 30, 82

[18] Mourão, L.; Couto, D. U. C.: *A fiscalização dos processos licitatórios na Administração Pública*. Revista do Tribunal de Contas do Estado de Minas Gerais, 2011. 12, 13

[19] Santos, F B e K R De Souza: *Como Combater a Corrupção em Licitações: Detecção e Prevenção de Fraudes*. FORUM, Belo Horizonte, 1ª edição, ISBN 9788545001652. `https://books.google.com.br/books?id=4i5qswEACAAJ`. 12, 13

[20] TRIBUNAL, DE CONTAS D A UNIÃO: *Glossário de Termos do Controle Externo*, 2017. 13

[21] BRASIL TRIBUNAL DE CONTAS DA UNIÃO, TCU: *Acórdão 1793/2011*, 2011. `http://www.tcu.gov.br/Consultas/Juris/Docs/judoc/Acord/20110801/AC_1793_27_11_P.doc`. 13

[22] Amaral, F.: *Aprenda Mineração de Dados: teoria e prática.* Alta Books, Rio de Janeiro, 1ª edição, 2016. 15

[23] Russell, Stuart e Peter Norvig: *Artificial Intelligence: A Modern Approach (2nd Edition)*. 2002, ISBN 0137903952. 15

[24] LUGER, G. F: *Inteligência Artificial*. PEARSON BRASIL, 6ª edição, 2013, ISBN 9788581435503. 15

[25] James, Gareth, Daniela Witten, Trevor Hastie e Robert Tibshirani: *An introduction to statistical learning*, volume 112. Springer, 2013. 15, 17, 18, 21, 22, 25

[26] Hastie, Trevor, Robert Tibshirani, Jerome Friedman e James Franklin: *The elements of statistical learning: data mining, inference and prediction*. The Mathematical Intelligencer, 27(2):83–85, 2005. 15, 16, 21, 22, 23, 24

[27] Friedman, J H: *Stochastic gradient boosting.* COMPUTATIONAL STATISTICS & DATA ANALYSIS, 38(4):367–378, fevereiro 2002, ISSN 0167-9473. 16

[28] Guelman, Leo: *Gradient boosting trees for auto insurance loss cost modeling and prediction.* EXPERT SYSTEMS WITH APPLICATIONS, 39(3):3659–3667, fevereiro 2012, ISSN 0957-4174. 16, 17

[29] Rao, Haidi, Xianzhang Shi, Ahoussou Kouassi Rodrigue, Juanjuan Feng, Yingchun Xia, Mohamed Elhoseny, Xiaohui Yuan e Lichuan Gu: *Feature selection based on artificial bee colony and gradient boosting decision tree.* Applied Soft Computing Journal, 2019, ISSN 15684946. 16

[30] Bhattacharyya, Siddhartha, Sanjeev Jha, Kurian Tharakunnel e J. Christopher Westland: *Data mining for credit card fraud: A comparative study.* Decision Support Systems, 50(3, SI):602–613, fevereiro 2011, ISSN 01679236. 17

[31] Breiman, L: *Random forests.* MACHINE LEARNING, 45(1):5–32, outubro 2001, ISSN 0885-6125. 17

[32] Cortes, Corinna e Vladimir Vapnik: *Support-Vector Networks.* Machine Learning, 1995, ISSN 15730565. 17

[33] Liu, Chuan, Wenyong Wang, Meng Wang, Fengmao Lv e Martin Konan: *An efficient instance selection algorithm to reconstruct training set for support vector machine.* Knowledge-Based Systems, 2017, ISSN 09507051. 17

[34] Drummond, Chris e Robert C Holte: *C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling.* Em *Workshop on learning from imbalanced datasets II*, 2003. 20

[35] Guo, Xinjian, Yilong Yin, Cailing Dong, Gongping Yang e Guangtong Zhou: *On the class imbalance problem.* Em *Proceedings - 4th International Conference on Natural Computation, ICNC 2008*, 2008, ISBN 9780769533049. 20, 21

[36] He, Haibo e Edwardo A. Garcia: *Learning from imbalanced data.* IEEE Transactions on Knowledge and Data Engineering, 2009, ISSN 10414347. 21, 26

[37] Rizzo, Maria L.: *Statistical Computing with R.* 2007. 21

[38] Kohavi, Ron: *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.* International Joint Conference of Artificial Intelligence, 1995. 21, 22

[39] Ebtehaj, Mohammad, Hamid Moradkhani e Hoshin V. Gupta: *Improving robustness of hydrologic parameter estimation by the use of moving block bootstrap resampling.* Water Resources Research, 2010, ISSN 00431397. 22

[40] Iguyon, Isabelle e André Elisseeff: *An introduction to variable and feature selection*, 2003. ISSN 15324435. 22, 23

[41] Carneiro, Nuno, Gonçalo Figueira e Miguel Costa: *A data mining based system for credit-card fraud detection in e-tail.* Decision Support Systems, 2017, ISSN 01679236. 22, 27

[42] Coussement, Kristof, Stefan Lessmann e Geert Verstraeten: *A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry.* Decision Support Systems, 2017, ISSN 01679236. 22, 27, 29

[43] Morettin, Pedro Alberto e WILTON OLIVEIRA BUSSAB: *Estatistica Basica.* Saraiva Educação SA, Sao Paulo, 6ª edição, 2010. 23

[44] Carvalho, Rommel N e Ricardo Silva Carvalho: *Bayesian Models to Assess Risk of Corruption of Federal Management Units.* Proceedings of the 13th UAI Bayesian Modeling Applications Workshop, 2016. 23, 28

[45] Nykodym, Tomas, Tom Kraljevic, Amy Wang e Wendy Wong: *Generalized linear modeling with H2O.* Published by H2O. ai Inc, 2016. 24

[46] Guyon, Isabelle, Jason Weston, Stephen Barnhill e Vladimir Vapnik: *Gene selection for cancer classification using support vector machines.* Machine Learning, 2002, ISSN 08856125. 24, 30

[47] Efron, Bradley, Trevor Hastie, Iain Johnstone, Robert Tibshirani, Hemant Ishwaran, Keith Knight, Jean Michel Loubes, Pascal Massart, David Madigan, Greg Ridgeway, Saharon Rosset, J. I. Zhu, Robert A. Stine, Berwin A. Turlach, Sanford Weisberg, Iain Johnstone e Robert Tibshirani: *Least angle regression.* Annals of Statistics, 2004, ISSN 00905364. 24

[48] Blatman, Géraud e Bruno Sudret: *Adaptive sparse polynomial chaos expansion based on least angle regression.* Journal of Computational Physics, 2011, ISSN 10902716. 24

[49] Wold, Svante, Kim Esbensen e Paul Geladi: *Principal component analysis.* Chemometrics and Intelligent Laboratory Systems, 1987, ISSN 01697439. 25

[50] Uğuz, Harun: *A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm.* Knowledge-Based Systems, 2011, ISSN 09507051. 25

[51] Tripathy, Abinash, Ankit Agrawal e Santanu Kumar Rath: *Classification of sentiment reviews using n-gram machine learning approach.* Expert Systems with Applications, 2016, ISSN 09574174. 25

[52] Smadi, Sami, Nauman Aslam e Li Zhang: *Detection of online phishing email using dynamic evolving neural network based on reinforcement learning.* Decision Support Systems, 2018, ISSN 01679236. 25, 26

[53] Ferri, C., J. Hernández-Orallo e R. Modroiu: *An experimental comparison of performance measures for classification.* Pattern Recognition Letters, 2009, ISSN 01678655. 25, 26

[54] Zhu, Wen, Nancy Zeng e Ning Wang: *Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations.* Northeast SAS Users Group 2010: Health Care and Life Sciences, 2010. 25, 26

[55] Fawcett, Tom: *An introduction to ROC analysis.* Pattern Recognition Letters, 2006, ISSN 01678655. 26

[56] Liu, Yangguang, Yangming Zhou, Shiting Wen e Chaogang Tang: *A Strategy on Selecting Performance Metrics for Classifier Evaluation.* International Journal of Mobile Computing and Multimedia Communications, 2014, ISSN 19379404. 26

[57] López-Iturriaga, Félix J. e Iván Pastor Sanz: *Predicting Public Corruption with Neural Networks: An Analysis of Spanish Provinces.* Social Indicators Research, 2018, ISSN 15730921. 27

[58] Yulianto, Aries, Adiwijaya, Moch Arif Bijaksana e Kemas M. Lhaksmana: *Fraud detection on international direct dial call using hybrid NBTree algorithm and Kullback Leibler divergence.* Em *2017 5th International Conference on Information and Communication Technology, ICoIC7 2017*, 2017, ISBN 9781509049127. 27

[59] Paula, Ebberth L., Marcelo Ladeira, Rommel N. Carvalho e Thiago Marzagão: *Deep learning anomaly detection as suppor fraud investigation in Brazilian exports and anti-money laundering.* Em *Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016*, 2017, ISBN 9781509061662. 27, 29

[60] Rizki, Adila Afifah, Isti Surjandari e Reggia Aldiana Wayasti: *Data mining application to detect financial fraud in Indonesia's public companies.* Em *Proceeding - 2017 3rd International Conference on Science in Information Technology: Theory and Application of IT for Education, Industry and Society in Big Data Era, ICSITech 2017*, 2018, ISBN 9781509058662. 27

[61] Moepya, Stephen Obakeng, Sharat Saurabh Akhoury, Fulufhelo Vincent Nelwamondo e Bhekisipho Twala: *The role of imputation in detecting fraudulent financial reporting.* International Journal of Innovative Computing, Information and Control, 2016, ISSN 13494198. 27

[62] Yee, O S, S Sagadevan e N.H.A.H. Malim: *Credit card fraud detection using machine learning as data mining technique.* Journal of Telecommunication, Electronic and Computer Engineering, 2018, ISSN 21801843 (ISSN). 27

[63] Mahmud, Mohammad Sultan: *An evaluation of computational intelligence in credit card fraud detection.* Em *20th International Computer Science and Engineering Conference: Smart Ubiquitos Computing and Knowledge, ICSEC 2016*, 2017, ISBN 9781509044207. 27

[64] Charleonnan, Anusorn: *Credit card fraud detection using RUS and MRN algorithms.* Em *2016 Management and Innovation Technology International Conference, MITiCON 2016*, 2017, ISBN 9781509041053. 27

[65] Monamo, Patrick M., Vukosi Marivate e Bhekisipho Twala: *A multifaceted approach to Bitcoin fraud detection: Global and local outliers*. Em *Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016*, 2017, ISBN 9781509061662. 27

[66] Li, Xurui, Wei Yu, Tianyu Luwang, Jianbin Zheng, Xuetao Qiu, Jintao Zhao, Lei Xia e Yujiao Li: *Transaction Fraud Detection Using GRU-centered Sandwich-structured Model*. Em *Proceedings of the 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design, CSCWD 2018*, 2018, ISBN 9781538614822. 27

[67] Peng, Hao e Mengzhuo You: *The health care fraud detection using the pharmacopoeia spectrum tree and neural network analytic contribution Hierarchy process*. Em *Proceedings - 15th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 10th IEEE International Conference on Big Data Science and Engineering and 14th IEEE International Symposium on Parallel and Distributed Proce*, 2016, ISBN 9781509032051. 27

[68] Zhou, Ligang, Yain Whar Si e Hamido Fujita: *Predicting the listing statuses of Chinese-listed companies using decision trees combined with an improved filter feature selection method*. Knowledge-Based Systems, 2017, ISSN 09507051. 27

[69] Taha, A A e S J Malebary: *An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine*. IEEE Access, 8:25579–25587, 2020, ISSN 2169-3536. 27

[70] Sapozhnikova, M. U., A. V. Nikonov, A. M. Vulfin, M. M. Gayanova, K. V. Mironov e D. V. Kurennov: *Anti-fraud system on the basis of data mining technologies*. Em *2017 IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2017*, 2018, ISBN 9781538646625. 27

[71] Awoyemi, John O., Adebayo O. Adetunmbi e Samuel A. Oluwadare: *Credit card fraud detection using machine learning techniques: A comparative analysis*. Em *Proceedings of the IEEE International Conference on Computing, Networking and Informatics, ICCNI 2017*, 2017, ISBN 9781509046423. 27

[72] Chouiekh, Alae e El Hassane Ibn El Haj: *ConvNets for fraud detection analysis*. Em *Procedia Computer Science*, 2018. 27

[73] Gupta, R Y, S Sai Mudigonda, P K Kandala e P K Baruah: *Implementation of a Predictive Model for Fraud Detection in Motor Insurance using Gradient Boosting Method and Validation with Actuarial Models*. Em *2019 IEEE International Conference on Clean Energy and Energy Efficient Electronics Circuit for Sustainable Development (INCCES)*, páginas 1–6, dezembro 2019. 27

[74] Rad, Mehdi Samee e Asadollah Shahbahrami: *Detecting high risk taxpayers using data mining techniques*. Em *Proceedings - 2016 2nd International Conference of Signal Processing and Intelligent Systems, ICSPIS 2016*, 2017, ISBN 9781509058204. 27

[75] Balaniuk, Remis, Pierre Bessiere, Emmanuel Mazer e Paulo Roberto Cobbe: *Corruption risk analysis using semi-supervised naïve Bayes classifiers*. International Journal of Reasoning-based Intelligent Systems, 2014, ISSN 1755-0556. 27

[76] Yaram, Suresh: *Machine learning algorithms for document clustering and fraud detection*. Em *Proceedings of the 2016 International Conference on Data Science and Engineering, ICDSE 2016*, 2017, ISBN 9781509012800. 27

[77] Silva, Vinicius Sarmento; Ralha, Celia Ghedini: *Utilização de Técnicas de Mineração de Dados como Auxílio na Detecção de Cartéis em Licitações*. Revista Eletrônica de Sistemas de Informação, 1, 2010. 27, 28

[78] Carvalho, Rommel N., Leonardo J. Sales, Henrique A. Da Rocha e Gilson L. Mendes: *Using Bayesian networks to identify and prevent split purchases in Brazil*. Em *CEUR Workshop Proceedings*, 2014. 28

[79] Balaniuk, Remis, Pierre Bessiere, Emmanuel Mazer e Paulo Cobbe: *Risk based government audit planning using na??ve bayes classifiers*. Frontiers in Artificial Intelligence and Applications, 2012, ISSN 09226389. 29

[80] Domingos, Silvio L., Rommel N. Carvalho, Ricardo S. Carvalho e Guilherme N. Ramos: *Identifying it purchases anomalies in the Brazilian Government Procurement System using deep learning*. Em *Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016*, 2017, ISBN 9781509061662. 29

[81] Zhang, Le Bing, Fei Peng, Le Qin e Min Long: *Face spoofing detection based on color texture Markov feature and support vector machine recursive feature elimination*. Journal of Visual Communication and Image Representation, 2018, ISSN 10959076. 30