UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS JURÍDICAS E POLÍTICAS
ESCOLA DE ADMINISTRAÇÃO PÚBLICA

GABRIEL ALVES DE FARIA

FRAUDES EM COMPRAS GOVERNAMENTAIS:
Detecção com Aprendizado de Máquina

Rio de Janeiro
2023

GABRIEL ALVES DE FARIA

FRAUDES EM COMPRAS GOVERNAMENTAIS:
Detecção com Aprendizado de Máquina

Trabalho de conclusão de curso apresentado à Escola de Administração Pública da Universidade Federal do Estado do Rio de Janeiro como requisito parcial para obtenção do grau de Bacharel em Administração Pública.

Orientador: Prof. Dr. Steven Dutt-Ross

Rio de Janeiro

2023

Faria, Gabriel Alves de
FA474f Fraudes em Compras Governamentais: Detecção com Aprendizado de Máquina / Gabriel Alves de Faria. -- Rio de Janeiro, 2023.
70

Orientador: Steven Dutt-Ross.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal do Estado do Rio de Janeiro, Graduação em Administração Pública, 2023.

1. Fraude. 2. Aprendizado de Máquina. 3. Compras governamentais. 4. Licitações. 5. Administração Pública. I. Dutt-Ross, Steven, orient. II. Título.

GABRIEL ALVES DE FARIA

FRAUDES EM COMPRAS GOVERNAMENTAIS:
Detecção com Aprendizado de Máquina

Trabalho de conclusão de curso apresentado à Escola de Administração Pública da Universidade Federal do Estado do Rio de Janeiro como requisito parcial para obtenção do grau de Bacharel em Administração Pública.

Aprovado em: 06 de fevereiro de 2023.

Banca examinadora:

_____

Prof. Dr. Steven Dutt-Ross (Orientador)

Universidade Federal do Estado do Rio de Janeiro (UNIRIO)

_____

Prof. Dr. Antônio Rodrigues de Andrade

Universidade Federal do Estado do Rio de Janeiro (UNIRIO)

_____

Prof.ª Dr.ª Letícia Martins Raposo

Universidade Federal do Estado do Rio de Janeiro (UNIRIO)

AGRADECIMENTOS

A Deus pelo cuidado e bençãos concedidas, dedico este trabalho e minha vida a Ele.

Aos meus pais, Suely Alves de Faria e Cláudio Luís Tavares de Faria, meus maiores exemplos de amor e carinho. Este trabalho não seria possível sem a dedicação de vocês, jamais poderei expressar toda minha gratidão, espero honrá-los e retribuir tamanho amor. Do fundo do meu coração, eu amo vocês.

À minha avó Vanilde Fontes de Souza Alves e meu avô José Alves Netto, tenho o privilégio de tê-los como verdadeiros pais ao meu lado, eu amo muito a senhora e o senhor.

Ao meu orientador, Prof. Dr. Steven Dutt-Ross, um educador singular e inspirador, agradeço a paciência ao longo dos anos para auxiliar no desenvolvimento desta pesquisa.

À Escola de Administração Pública da Universidade Federal do Estado do Rio de Janeiro, por tornar possível o sonho de me tornar bacharel. Obrigado a todos os professores por fazerem parte da minha formação acadêmica.

RESUMO

Com o advento da disponibilização de grandes bases de dados por parte do Governo Federal à sociedade, em consideração ao princípio da transparência, diversos estudos tornam-se possíveis, como, por exemplo, a identificação de fraudes em processos de compras governamentais por meio de aplicações de inteligência artificial. O fenômeno das fraudes lida com eventos minuciosos. Na esteira do avanço de tecnologias computacionais, o combate à corrupção pode valer-se do uso de técnicas como o aprendizado de máquina para detectar padrões e prever possíveis ocorrências de fraudes em operações de compras públicas. O objetivo geral desta pesquisa foi investigar a detecção de fraudes em compras governamentais no Brasil por meio técnicas de aprendizado de máquina. A pesquisa buscou identificar quais técnicas são mais eficazes, com o intuito de contribuir para a melhoria das medidas de prevenção e combate à fraude no setor público brasileiro. Após a revisão bibliográfica de temas como compras governamentais, fraude à licitação e aprendizado de máquina, foram utilizados seis algoritmos preditivos: Regressão Logística, *Random Forest*, Redes Neurais, *Naïve Bayes*, *Stochastic Gradient Boosting* e Árvore de Decisão. Do ponto de vista teórico e prático, este trabalho ajuda a entender se o emprego de modelos preditivos de classificação pode auxiliar o gestor público na identificação de empresas fraudadoras. Os resultados demonstraram a inexigibilidade de licitação, a escolha da modalidade pregão e o valor do contrato como fatores importantes para ocorrência de fraudes, o que sugere a possibilidade de mudanças em processos governamentais. Por fim, são apresentadas sugestões para futuras pesquisas.

Palavras-chave: Fraude; Aprendizado de Máquina; Compras Governamentais; Licitações; Administração Pública.

ABSTRACT

With the advent of the availability of large databases by the Federal Government to society, in consideration of the principle of transparency, various studies become possible, such as, for example, the identification of frauds in government procurement processes through artificial intelligence applications. The phenomenon of frauds deals with meticulous events. In the wake of the advance of computational technologies, the fight against corruption can make use of techniques such as machine learning to detect patterns and predict possible occurrences of frauds in public procurement operations. The overall objective of this research was to investigate the detection of fraud in government procurement in Brazil through machine learning techniques. The research sought to identify which techniques are most effective, with the aim of contributing to the improvement of fraud prevention and combat measures in the Brazilian public sector. After a literature review of topics such as government procurement, bid rigging and machine learning, six predictive algorithms were used: Logistic Regression, Random Forest, Neural Networks, Naive Bayes, Stochastic Gradient Boosting and Decision Tree. From a theoretical and practical point of view, this work helps to understand if the use of classification predictive models can assist the public manager in identifying fraudulent companies. The results showed the unenforceability of bidding, the choice of the bidding mode and the value of the contract as important factors for the occurrence of frauds, suggesting the possibility of changes in government processes. Finally, suggestions for future research are presented.

Keywords: Fraud; Machine Learning; Government Procurement; Bidding; Public Administration.

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
ASM	Medida de Seleção de Atributos
BNDES	Banco Nacional de Desenvolvimento Econômico e Social
CART	*Classification and Regression Tree*
CEIS	Cadastro Nacional de Empresas Inidôneas e Suspensas
CGU	Controladoria Geral da União
CNPJ	Cadastro Nacional de Pessoas Jurídicas
CSV	*Comma-separated Values*
EIS	Empresa Inidônea ou Suspensa
EM	Expectativa-Maximização
EPL	Empresa de Planejamento e Logística S.A.
IA	Inteligência Artificial
MDP	*Markov Decision Process*
ME	Ministério da Economia
PAC	Programa de Aceleração do Crescimento
PPI	Programa de Parceria de Investimento
RDC	Regime Diferenciado de Compras Governamentais
RFP	*Request For Proposal*
RNAs	Redes Neurais Artificiais
ROSE	*Random Over-Sampling Examples*
TCU	Tribunal de Contas da União
TI	Tecnologia da Informação

SUMÁRIO

1 INTRODUÇÃO

1.1 APRESENTAÇÃO DO TEMA

O presente trabalho de conclusão de curso tem o objetivo de estudar a aplicação de técnicas de aprendizado de máquina ne detecção de fraudes no processo licitatório das compras públicas. É cada vez mais comum o uso de tecnologias de detecção de fraudes nas operações de compras públicas para garantir a correta aplicação dos recursos públicos. Essas tecnologias, quando usadas corretamente, oferecem avanços significativos para prevenção e investigação de comportamentos irregulares. Tais técnicas são capazes de envolver ferramentas computacionais para buscar padrões em grandes volumes de dados e identificar potenciais fraudes.

A detecção de fraudes em compras públicas é uma das mais complexas áreas para se trabalhar devido à grande variedade de operações envolvidas. Nesse contexto, a análise de fraudes se torna ainda mais importante para assegurar que as operações realizadas cumpram as diretrizes de legalidade. Empregando técnicas de Inteligência, as informações fornecidas por auditorias financeiras regulares podem ser analisadas automaticamente, minimizando assim os riscos inerentes à tomada de decisões inadequadas.

O uso de algoritmos de aprendizado de máquina, subcampo da Inteligência Artificial, para detectar fraudes em compras públicas pode ser de grande ajuda para as entidades governamentais locais que sofrem atos fraudulentos. Isso ocorre, pois, quando aplicadas, essas técnicas reduzem significativamente o custo envolvido com a auditoria e monitoramento de compras.

1.2 OBJETIVO GERAL

O presente estudo teve como objetivo principal investigar a detecção de fraudes em compras governamentais no Brasil por meio técnicas de aprendizado de máquina. A pesquisa buscou identificar quais técnicas são mais eficazes na detecção de fraudes, com o intuito de contribuir para a melhoria das medidas de prevenção e combate à fraude no setor público brasileiro.

1.3 OBJETIVOS ESPECÍFICOS

Para responder ao problema da pesquisa, foram formulados os seguintes objetivos específicos:

a) investigar e definir o conceito de fraudes nas compras públicas no Brasil;

b) desenvolver modelos de previsão de fraudes em compras governamentais no Brasil utilizando técnicas de aprendizado de máquina, como Regressão Logística, *Random Forest*, Redes Neurais, *Naïve Bayes* e *Stochastic Gradient Boosting* e Árvore de Decisão.

c) avaliar o desempenho dos modelos desenvolvidos.

d) utilizar os modelos criados para previsão de empresas potencialmente sujeitas à aplicação da sanção de suspensão do direito de licitar e declaração de inidoneidade.

1.4 PERGUNTAS DA PESQUISA

Para atingir os objetivos específicos estabelecidos, foram determinadas as seguintes questões:

a) quais são os desafios enfrentados na detecção de fraudes em compras governamentais e como as técnicas de aprendizado de máquina podem ser usadas para superá-los?

b) quais técnicas de aprendizado de máquina são mais eficazes na detecção de fraudes em compras governamentais no Brasil?

c) quais são as variáveis explicativas mais importantes para a variável resposta “empresas inidôneas e suspensas por fraudes em compras governamentais no Brasil”?

d) quais são as implicações dos resultados da pesquisa para a gestão pública e a prevenção de fraudes em compras governamentais no Brasil?

1.5 JUSTIFICATIVAS

A presente pesquisa tem como objetivo investigar e avaliar o desempenho de técnicas de aprendizado de máquina na detecção de fraudes em compras governamentais no Brasil. Isso se justifica devido ao fato de fraudes em compras governamentais são um problema grave no país, causando prejuízos financeiros e prejudicando a confiança pública. Utilizando técnicas de aprendizado de máquina, pretende-se desenvolver modelos eficazes para prevenir e detectar fraudes, possibilitando selecionar empresas para investigação. Desta forma, a pesquisa busca contribuir para o avanço do conhecimento e fornecer soluções práticas para o problema das fraudes em compras governamentais.

1.6 DELIMITAÇÃO DO ESTUDO

Este estudo se concentrará nos registros de contratos de compras governamentais na esfera federal realizadas no Brasil entre 1990 e 2022, e informações sobre empresas declaradas inidôneas ou suspensas pelo Poder Público de 1988 até 2022. Ressalta-se que foram desconsideradas observações sobre pessoas físicas, pois envolvem dados pessoais potencialmente sensíveis. Foram empregadas técnicas de aprendizado de máquina para detectar fraudes, como modelos de classificação, não sendo aplicáveis ao caso concreto, os modelos de regressão.

2 REFERENCIAL TEÓRICO

2.1 COMPRAS GOVERNAMENTAIS

A Constituição brasileira estabelece que, regra geral, todas as compras e vendas realizadas e todos os serviços e obras contratados pela administração pública devem ser precedidos de uma licitação pública. A Lei Federal nº 8.666/1993, que deve ser observada pelos três ramos do governo, estabelece a estrutura geral aplicável a todas as licitações públicas no país (FONSECA, 2014).

A Constituição também estabelece princípios pelos quais a administração pública brasileira está vinculada: legalidade, impessoalidade, moralidade, publicidade e eficiência. A Lei Federal nº 8.666/1993 acrescentou a eles os conceitos de aderência estrita à solicitação de proposta ou *request for proposal* (RFP) e julgamento objetivo como princípios que devem ser observados em licitações públicas. Como consequência, o licitante vencedor não é a única entidade legalmente vinculada aos termos e condições estabelecidos na RFP (SANTANNA e DANIEL, 2016, p. 3).

Estatutos diferentes regulam licitações e contratos públicos específicos. A Lei Federal nº 8.987/1995, por exemplo, estabelece a estrutura geral aplicável aos serviços e concessões e permissões de obras públicas. Como as concessões geralmente envolvem contratos de longo prazo e os serviços públicos são de extrema relevância para a economia brasileira e para o bem-estar de seus cidadãos, esse estatuto continua sendo de grande importância no arcabouço jurídico local (FONSECA, 2014).

Em 2004, a Lei Federal nº 11.079/2004 introduziu parcerias público-privadas (PPPs) no sistema jurídico brasileiro. Na sua versão nacional, consistem basicamente em uma concessão de serviços públicos ou obras públicas, na qual a compensação a ser paga à parte privada resultará de uma combinação entre tarifas cobradas dos cidadãos e pagamentos diretos feitos pela administração (SANTANNA e DANIEL, 2016, p. 3).

Em 2011, o Congresso aprovou a Lei Federal nº 12.462/2011, que criou o Regime Diferenciado de Compras Governamentais (RDC). Esse regime foi criado para agilizar os procedimentos de licitação das obras de infraestrutura necessárias para o Brasil sediar a Copa do Mundo da FIFA 2014 e os Jogos Olímpicos de 2016 no Rio de Janeiro. Posteriormente, no entanto, esse estatuto foi sucessivamente alterado para permitir a sua aplicação a outros projetos públicos, como:

- obras de infraestrutura incluídas no Programa de Aceleração do Crescimento (PAC) criado pelo governo federal;

- obras e serviços de engenharia relacionados à saúde pública;

- obras e serviços de engenharia para a construção e reforma de instalações criminosas;

- ações em segurança pública;

- obras e serviços de engenharia relacionados à mobilidade urbana e infraestrutura logística.

Em 2016, o Projeto de Lei das Empresas Estatais (Lei Federal nº 13.303/2016) criou uma estrutura específica para licitações públicas realizadas por empresas estatais que deveria ser mais flexível do que as regras gerais com base no objetivo de fazer as empresas estatais competirem no mercado. No mesmo ano, foi lançado um novo programa governamental chamado Programa de Parceria de Investimento (PPI) para coordenar os projetos de infraestrutura mais importantes em nível nacional e para servir como balcão único para investidores e outras partes interessadas (BARROSO e BARROSO, 2017).

O PPI é composto por um conselho, chefiado pelo próprio Presidente e formado por ministérios centrais e setoriais, junto com os presidentes do Banco Nacional de Desenvolvimento Econômico e Social (BNDES), Caixa Econômica Federal e Banco do Brasil, e uma secretaria, com pessoal para exercer funções consultivas no conselho, com o apoio da Empresa de Planejamento e Logística S.A. (EPL).

Em 2021, o novo regramento sobre Licitações e Contratos Administrativos foi instituído pela Lei nº 14.133/2021 e trouxe uma série de inovações, tais como a exclusão das modalidades de carta-convite e tomada de preços e a inclusão de uma nova modalidade: o diálogo competitivo. A nova lei também estabelece que os processos de licitação devem ocorrer preferencialmente por meios digitais (art. 12, inciso VI). As licitações presenciais são consideradas exceção e devem ser justificadas e ter as sessões obrigatoriamente registradas em ata e gravadas em áudio e vídeo.

Desde a data de sua publicação oficial, a Lei nº 14.133/2021 revogou os artigos 89 a 108 da Lei 8.666/1993 (art.193, inciso I) e entrou em vigor (art.194, *caput*). No entanto, foi

estabelecido um prazo de dois anos, a partir da publicação, para a transição e ab-rogação das Leis 8.666/1993, 10.520/2002 e 12.462/2011 (artigo 193, inciso II).

2.1.1 Licitações públicas no Brasil

O artigo 2º da Lei nº 8.666, de 21 de junho de 1993, dispõe:

> Art. 2º. As obras, serviços, inclusive anúncios, aquisições, alienações, concessões, alvarás e arrendamentos da Administração Pública, quando contratados com terceiros, deverão ser precedidos de procedimento licitatório, ressalvadas as hipóteses previstas nesta lei.
> Parágrafo único. Para efeitos desta Lei, considera-se contrato qualquer acordo entre organismos ou entidades da Administração Pública e pessoas que inclua um acordo de vontades para a constituição de uma caução e a prestação de obrigações recíprocas, independentemente da moeda utilizada (BRASIL, 1993, p. 1).

E ainda de acordo com o artigo 3º da mesma lei, a licitação tem como objetivo selecionar a proposta mais vantajosa para a gestão pública dentro dos "princípios fundamentais da legalidade, impessoalidade, moralidade, igualdade, publicidade, probidade, procedimento administrativo, o vínculo com a citação, o julgamento objetivo e os relacionados a eles".

O processo licitatório visa escolher a alternativa que oferece maior qualidade e menores preços para a realização das atividades, garantindo que o contratado cumpra com as especificações escritas (MIRANDA, 2010).

Antes de realizar uma licitação, a Administração deve avaliar a necessidade do serviço, seus benefícios e a economia da medida, considerando também a impossibilidade de sua implementação sem prejudicar o interesse público. (PINTO, 2020).

Se a Administração tiver condições de exercer a função diretamente, deve fazê-lo; caso contrário, a terceirização será considerada ilícita e os responsáveis estarão sujeitos às penalidades previstas em lei (MIRANDA, 2010).

Do mesmo modo, no que diz respeito à execução dos contratos celebrados pela Administração Pública, a impessoalidade deve ser sempre valorizada e um agente fiscalizador especialmente designado para esta função deve cumprir o dever de fiscalização. Este fiscal tem como objetivo observar a provisão para a prestação dos serviços e a alocação de recursos, com o intuito de garantir o cumprimento de todos os termos estipulados no contrato (artigo 67 da Lei nº 8.666/93 e Instrução Normativa nº 2/ 2008). Além disso, a Instrução Normativa nº 6, de 23 de dezembro de 2013, destaca que a revisão dos contratos deve se basear em critérios estatísticos, levando em conta não apenas falhas, mas também deficiências que afetam o contrato como um todo.

Desta forma, a Sumula nº 331 (BRASIL, 2011) inclui especificações sobre o contrato a ser celebrado entre o empreiteiro e o empreitante (empresa e Administração Pública) caso não haja vínculo empregatício independente do setor de atividade da empresa contratada, Lei n. 8.666/93 deve ser observada durante a prestação do serviço, entre outras.

Para a contratação de empresas do setor privado para prestação de serviço terceirizado ou aquisição de bens, a íntegra do procedimento licitatório deve ser tornada pública (OLIVEIRA, 2019). Licitação é uma oferta (normalmente competitiva) para estabelecer um preço para um produto ou serviço por uma pessoa ou empresa, ou uma solicitação para que algo seja feito. O lance determina o preço ou o valor de algo (ROSILHO, 2011).

Dependendo das circunstâncias, um "comprador" ou "fornecedor" de um produto ou serviço pode apresentar uma oferta. Em leilões, bolsas de valores e imóveis, a quantia que uma empresa ou pessoa está disposta a pagar é chamada de oferta (OLIVEIRA, 2019). No contexto de compras comerciais ou governamentais, a oferta de preço pela qual uma empresa ou indivíduo está disposto a vender é às vezes chamada de oferta.

No mundo tecnologicamente sofisticado de hoje, a Internet é a plataforma ideal para fornecer serviços de licitação; a licitação é um método natural para estabelecer o preço de um produto em uma economia de livre mercado (RUTZ, 2018).

O processo licitatório é um procedimento administrativo, isonômico, no qual a administração examina e escolhe a proposta mais vantajosa para compra de produtos ou para registro de preços para contratos futuros. Nenhum processo licitatório poderá ser realizado em segredo; deve ser sempre transparente e aberto a todos os cidadãos e empresas (PINTO, 2020).

As compras governamentais e compras públicas no Brasil são agora controladas pela regulamentação 8.666/93 de regras básicas para licitações e contratos e pela Lei 10.520/02, geralmente chamada de lei comercial. Segundo o Tribunal de Contas da União (TCU), também existe uma lei complementar na forma das Leis 8.666/93 de normas gerais de licitações e contratos, Lei 10.520/02 do Regime Diferenciado de Contratações Públicas e Decreto 5.450/2005 do Formulário de Leilão Eletrônico. Assim, as licitações públicas contribuem para a oferta de competitividade e crescimento social do mercado econômico, com o objetivo de melhorar a arrecadação de impostos, a arrecadação e a taxa de desemprego.

Segundo Motta (2011), na legislação brasileira, a oferta mais vantajosa é aquela que atende aos critérios do menor preço, da melhor técnica, melhor técnica e preço, ou do maior lance ou oferta em caso de alienação de bens ou concessão de direito genuíno de uso. O critério “menor preço” é frequentemente empregado entre eles. Existem duas ordens de categorização para os tipos de licitações que devem ser apresentadas no Brasil.

2.1.2 Os princípios basilares aplicados às licitações

O objetivo da contratação pública é conceder contratos de maneira oportuna e econômica a empreiteiros, fornecedores e prestadores de serviços qualificados para o fornecimento de bens, obras e serviços que apoiem as operações do governo e do serviço público, de acordo com os princípios e procedimentos estabelecido pelas regras de contratação pública (ROSILHO, 2011).

A contratação pública é baseada nos princípios de contratação pública, que devem ser abordados nos regulamentos de contratação pública. Eles fornecem a base para um código de conduta para profissionais de contratação pública e qualquer outro funcionário envolvido direta ou indiretamente com o processo de contratação pública (TEIXEIRA, 2011). Transparência, honestidade, economia, abertura, justiça, competitividade e prestação de contas são características básicas da contratação pública:

- Transparência: todas as partes interessadas na contratação pública, incluindo empreiteiros, fornecedores e prestadores de serviços, devem ter acesso a informações sobre o processo de contratação pública (GUIMARÃES, 2011);

- Integridade: a integridade nos contratos públicos é dupla, abrangendo tanto a do procedimento de contratação, quanto a dos especialistas em contratação pública (MIRANDA, 2010);

- Economia: o conceito de economia enfatiza a necessidade de gerir os fundos públicos com cuidado e atenção para garantir que os preços pagos pelos produtos, serviços e obras sejam aceitáveis e ofereçam um ótimo valor para os fundos de despesa pública (TEIXEIRA, 2011);

- Abertura: as normas para a contratação pública devem ser acessíveis a todas as empresas e pessoas qualificadas, e o público deve ter acesso aos requisitos para contratação pública (PEREIRA, 2010);

- Equidade: em vez de definir a justiça em contratos públicos como tratar todas as propostas igualmente, é preferível descrever como a justiça é alcançada em contratos públicos devido a interpretações variadas (GUIMARÃES, 2011);

- Concorrência: os requisitos de contratação pública devem ser amplamente divulgados para maximizar a probabilidade de uma reação favorável do mercado, resultando na adjudicação de contratos a preços competitivos (PEREIRA, 2010);

- Prestação de contas: implica que todos os participantes no processo de contratação pública sejam responsáveis por suas ações e escolhas (MIRANDA, 2010).

Não há nada que seja realmente local no domínio das compras governamentais. Embora uma instituição pública possa operar dentro de um setor específico de uma jurisdição, seu sistema operacional de compras é tipicamente influenciado por padrões globais formalmente estabelecidos e práticas de compras governamentais amplamente adotadas no setor público, que evoluíram para padrões globalmente reconhecidos (GUIMARÃES, 2011).

A base legal para a aquisição é frequentemente descrita em regras e procedimentos de aquisição e gestão de contratos, manuais e diretrizes, bem como em formulários de solicitação padrão usados para obter propostas de empreiteiros, fornecedores e prestadores de serviços (PINTO, 2020). A linguagem das regras, processos, diretrizes, manuais e documentos padrão que regem as compras públicas deve ser consistente com a estrutura legislativa que rege as compras públicas. Qualquer violação da legislação que rege os contratos públicos é ilegal e punida por lei.

2.1.3 Tipos de licitação

Inicialmente, é fundamental entender que "tipos de licitação" não implicam em "estilo de licitação". A Administração considera o "Tipo" ao decidir sobre a melhor proposta a seguir. Os fatores mais comuns usados para avaliar lances e propostas são o preço mais baixo, a melhor abordagem e o custo.

A abordagem para o melhor tipo de técnica, conforme citado por Barroso e Barroso (2017), está disposta no item 1, art. 46 da Lei nº 8.666/93. As propostas técnicas dos licitantes pré-aprovados serão revisadas em relação aos critérios de avaliação especificados e receberão pontos com base na qualidade do serviço prestado.

A Administração escolhe a proposta com o melhor custo-benefício e eficiência de tempo, com base na média ponderada dos componentes técnicos e financeiros das propostas apresentadas (CHENISK, 2008).

2.1.4 Modalidades de licitação

As modalidades de licitação vigentes no Brasil foram estabelecidas pela Lei nº 8.666/93, sendo elas:

- Concorrência - nesta modalidade, todos os interessados que comprovem preencher os requisitos descritos no edital podem participar. O anúncio de notificação pública será amplamente divulgado e os bens serão adquiridos por meio de licitação competitiva, com valores monetários altamente estimados;

- Convite - esta modalidade é realizada entre inscritos ou não, selecionados e convidados, sendo o número mínimo de convocados definido pela unidade administrativa igual a três. A data, local e documento de solicitação são definidos sem anúncio público. É importante entender que este modo é projetado para itens de baixo preço;

- Tomada de preços - as aquisições de mercadorias de valor intermediário são feitas por tomada de preços, que pode ser substituída por outras modalidades, como concurso ou convite, desde que os interessados estejam cadastrados e cumpram as exigências de notificação até o terceiro dia anterior ao dia do recebimento das ofertas;

- Concurso - as partes envolvidas podem ser qualquer pessoa com interesse nas artes, ciências ou empreendimentos técnicos e desejando conceder reconhecimento monetário ou outras formas de reconhecimento aos mais merecedores;

- Leilão - nesta modalidade, qualquer pessoa com interesse investido pode vender itens ou coisas que foram confiscadas ou penhoradas pelo governo, mas não são úteis para o Estado.

De acordo com Meirelles (2004), a dispensa de licitação só pode ser justificada por uma emergência reconhecida e declarada, com o objetivo de corrigir uma anormalidade ou evitar prejuízos. Em casos de guerra, perturbação grave da ordem ou calamidade pública, a dispensa de licitação pode ser autorizada na área afetada.

Uma sexta modalidade de licitação é o pregão eletrônico, regulado pela Lei nº 10.520/02, que conforme Bandeira de Mello (2009)

> O pregão eletrônico é uma modalidade de licitação que se realiza por meio da Internet, utilizando ferramentas eletrônicas para garantir a transparência e a eficiência dos processos. A principal vantagem do pregão eletrônico é a sua agilidade, pois permite que os lances sejam feitos de forma mais rápida e eficiente, sem a necessidade de reuniões presenciais. Além disso, o pregão eletrônico também oferece mais transparência ao processo de licitação, pois permite que todas as informações sejam registradas eletronicamente, garantindo maior segurança e eficiência na tomada de decisões (MELLO, 2019, p. 456).

Nesta modalidade, as propostas e lances são feitos em uma sessão aberta, presencial ou online, sendo selecionada a proposta ou oferta de maior e melhor valor para o contratante.

2.2 FRAUDES EM COMPRAS GOVERNAMENTAIS

Fraude é uma ação intencionalmente enganosa projetada para fornecer ao perpetrador um ganho ilegal ou para negar um direito à vítima. Os tipos de fraude incluem fraude fiscal, fraude de cartão de crédito, fraude eletrônica, fraude de valores mobiliários e fraude de falência (CAZELLA, 2019).

A fraude é um ato malicioso, desonesto e de má-fé com a intenção de prejudicar ou enganar outras pessoas para ganho financeiro do fraudador, muitas vezes às custas da vítima. De acordo com o artigo 171 do Código Penal Brasileiro, a fraude abrange um universo repleto de diversos delitos e punições. Pode causar danos financeiros, psicológicos e até mesmo à reputação da vítima que são irreparáveis.

Ela também envolve a representação falsa de fatos, seja pela retenção intencional de informações importantes ou pelo fornecimento de declarações falsas a outra parte, com o objetivo específico de obter algo que não poderia ter sido concedido sem o engano. Frequentemente, o perpetrador da fraude está ciente de informações que a vítima pretendida

não tem, permitindo que o perpetrador a engane. No fundo, o indivíduo ou empresa que comete fraude está se aproveitando da assimetria de informações (COSTA *et al.*, 2022).

Tanto os estados federativos quanto o governo federal têm leis que criminalizam a fraude, embora ações fraudulentas nem sempre resultem em julgamento criminal. Os promotores do governo muitas vezes têm poder discricionário substancial para determinar se um caso deve ir a julgamento e podem, em vez disso, buscar um acordo se isso resultar em uma resolução mais rápida e menos dispendiosa. Se um caso de fraude for a julgamento, o perpetrador pode ser considerado culpado e sentenciado à prisão (CAZELLA, 2019).

Embora o governo possa decidir que um caso de fraude pode ser resolvido fora do processo criminal, as partes não-governamentais que alegam danos podem prosseguir com um processo civil. As vítimas de fraude podem processar o perpetrador para recuperar os fundos ou, no caso em que não houve perda monetária, podem processar para restabelecer os direitos da vítima (SAMPAIO e FIGUEIREDO, 2019).

2.2.1 O contexto das fraudes em processos de compras públicas

A movimentação de bilhões de reais via licitações no Brasil tem impacto direto no PIB do país. Tendo em vista que a instituição tem caráter normativo constitucional, é inegável sua importância no ordenamento jurídico nacional: o artigo 37, inciso XXI, da Constituição de 1988 a estabelece como uma das normas fundamentais da Administração Pública. Apesar da alta prevalência de fraudes neste setor, conforme demonstrado pela atuação dos órgãos reguladores, a Lei nº 8.666, de 1993, foi alterada para oferecer maiores padrões de eficiência e moralidade às licitações e contratações administrativas.

O artigo 37, inciso XXI, da Constituição Federal estabelece, como regra geral, que os órgãos da Administração Pública somente poderão adquirir bens e serviços por meio de licitação. No entanto, o próprio artigo constitucional permite que a lei disciplinar pertinente renuncie ou julgue a licitação inexequível em determinadas circunstâncias. O artigo 24 da Lei nº 8.666/93 descreve as circunstâncias em que a licitação é dispensável.

Conforme determina a Constituição, todas as compras e contratos governamentais devem estar sujeitos a um procedimento de licitação. É um procedimento em que as empresas concorrem pela prestação de um serviço ou pela venda de um produto que a Administração Pública venha a utilizar. Devido a protocolos insuficientes, os processos licitatórios são suscetíveis a fraudes (ISHIKAWA e DE ALENCAR, 2020).

A fraude pode ocorrer em qualquer momento do processo licitatório, inclusive na primeira publicação do edital, durante a sessão de licitação e até mesmo após a aprovação ou quando da assinatura do contrato administrativo (COSTA *et al.*, 2022). Inquestionavelmente, as licitações devem seguir os princípios e padrões descritos na Lei nº 8.666/93.

> Art. 3° A licitação destina-se a garantir a observância do princípio constitucional da isonomia, a seleção da proposta mais vantajosa para a administração e a promoção do desenvolvimento nacional sustentável e será processada e julgada em estrita conformidade com os princípios básicos da legalidade, da impessoalidade, da moralidade, da igualdade, da publicidade, da probidade administrativa, da vinculação ao instrumento convocatório, do julgamento objetivo e dos que lhes são correlatos (BRASIL, 1993, p. 1).

Caso alguma dessas condições não seja cumprida, o processo licitatório será encerrado, podendo a licitação ser anulada, revogada ou julgada como administrativamente improvável. Além disso, a segmentação ocorre quando a agência e o licitante escolhido já estão trabalhando juntos para que o licitante escolhido vença a licitação. Desta forma, entende-se que não houve uma concorrência, uma vez que o vencedor da licitação já teria sido definido previamente (SAMPAIO e FIGUEIREDO, 2019).

Quando o edital é fraudulento, são inseridas cláusulas que cerceiam a participação de outros licitantes ou mesmo são inúteis para processo, não sendo requisitos para o objeto da licitação (CAZELLA, 2019).

O objeto da licitação deve ser individualizado e de acordo com as necessidades da Administração Pública. Fraudes também podem ocorrer em relação à descrição do ativo a ser adquirido ou à qualificação das empresas, dificultando a participação de muitas empresas e facilitando o sucesso de uma única empresa (ISHIKAWA e DE ALENCAR, 2020).

Além disso, um dos principais fatores a ser avaliado em um processo licitatório são os documentos necessários para a participação, tanto os relativos à qualificação dos licitantes quanto os necessários ao objeto da licitação. É comum nesta forma de fraude a existência de organizações “fantasmas”, aquelas que realmente não existem e falsificam documentos para ganhar a licitação e embolsar o dinheiro sem cumprir o serviço prometido (CAZELLA, 2019).

Lopes (2019, p. 23) explica que

> Na visão dos pregoeiros, os mecanismos de controle mais importantes para evitar a corrupção em pregões são a existência de um *check-list* e publicidade e propaganda.

> O checklist é uma lista de verificação para viabilizar a contratação por meio de diretrizes únicas e organizadas. O mecanismo de publicidade e propaganda é a obrigação da Administração Pública em divulgar, de maneira ampla, os editais de licitações na rede mundial de computadores (LOPES, 2019, p. 23).

Empresas que não possuem os documentos necessários para serem contratadas para o objeto da licitação falsificam os papéis para participar em muitos casos de licitações fraudulentas. As fraudes nas propostas incluem, por exemplo, a composição de preços, ganhos inflacionados e preços inexequíveis. Em seguida, a fraude de preço de licitação pode ocorrer de diversas formas, sendo uma delas o orçamento superfaturado, em que o orçamento fornecido pelo órgão licitante e o preço previsto do contrato já estão superfaturados, ou seja, com preços acima da realidade de mercado (ISHIKAWA e DE ALENCAR, 2020).

Quando empresas relacionadas utilizam as vantagens da lei para participar do processo de licitação, também pode ocorrer fraude. Isso pode ser observado, por exemplo, na aliança entre uma pequena e uma grande corporação, em que a empresa maior tenta aproveitar as vantagens que a lei oferece às microempresas. Além disso, a presença de vínculos entre os licitantes prejudica os princípios fundamentais de igualdade e competição em qualquer processo licitatório (COSTA *et al.*, 2022).

Outro tipo de fraude surge quando as Pessoas Jurídicas participantes do evento possuem um mesmo controlador. Isso permite que um mesmo indivíduo apresente várias ofertas em nome de várias empresas, comprometendo a confidencialidade e prejudicando a igualdade entre os licitantes. A formação de “cartéis de licitações” em aquisições, ou seja, acordos de conluio em licitações, é estimulada por dinâmicas muito semelhantes àquelas que controlam a formação de cartéis em mercados oligopolistas, vistos como uma aliança de corporações que buscam predeterminar o vencedor (CAZELLA, 2019).

Um cartel é um acordo de atores econômicos para restringir a concorrência. Esses cartéis são prejudiciais ao Estado porque aumentam os preços e limitam a disponibilidade de bens ou serviços, ou impossibilitam sua aquisição. Por isso, é fundamental que haja concorrência entre as licitações para que o licitante vencedor consiga o melhor preço para a administração pública, obtendo assim o máximo benefício (SAMPAIO e FIGUEIREDO, 2019).

No entanto, quando surgem os cartéis, os preços sobem e a concorrência é enfraquecida restando poucas organizações. Com o intuito de coibir cartéis, a Lei nº 12.529/11 regula a estrutura do Sistema Brasileiro de Defesa da Concorrência, dispõe sobre a prevenção e

repressão de infrações à ordem econômica e descreve as ações que implicam na formação de um cartel e acarretam sanções administrativas (COSTA *et al.*, 2022).

2.3 INTELIGÊNCIA ARTIFICIAL

A inteligência artificial (IA), muitas vezes conhecida como inteligência de máquina, é a inteligência mostrada por máquinas em oposição à inteligência natural exibida por humanos e outros animais. No uso comum, IA refere-se a computadores que imitam atividades "cognitivas" atribuídas a outros cérebros humanos, como "aprendizado" e "resolução de problemas" (SOUZA *et al.*, 2021).

Existem três tipos distintos de sistemas de inteligência artificial: inteligência artificial analítica, inteligência artificial inspirada em humanos e inteligência artificial humanizada. A IA analítica tem apenas traços de inteligência cognitiva, gera representações cognitivas do ambiente e usa o aprendizado baseado em experiências anteriores para influenciar julgamentos futuros (MOREIRA *et al.*, 2021).

Além da inteligência cognitiva, a IA inspirada em humanos entende e leva em consideração as emoções humanas em suas tomadas de decisão. A IA humanizada demonstra todos os tipos de inteligência (cognitiva, emocional e social) e é capaz de autoconhecimento e consciência nas interações sociais (MATIAS, 2020).

O campo acadêmico da inteligência artificial foi estabelecido em 1956, e os anos seguintes foram marcados por múltiplas ondas de entusiasmo, seguidas de decepção e falta de financiamento (conhecido como "inverno da IA"), novos métodos, sucesso e investimentos renovados. Durante a maior parte de sua existência, a pesquisa de IA foi dividida em subcampos que interagem entre si (FÉLIX, 2020).

Esses subcampos se distinguem por fatores tecnológicos, como objetivos específicos (por exemplo, "robótica" ou "aprendizagem de máquina"), o uso de certas ferramentas ("lógica") ou diferenças filosóficas profundas (POSTAL, 2019).

As questões tradicionais de pesquisa de IA incluem raciocínio, representação de conhecimento, planejamento, aprendizado, processamento de linguagem natural, percepção e manipulação de objetos. A inteligência geral é um dos objetivos de longo prazo do campo. Técnicas estatísticas, inteligência computacional e IA clássica estão entre as abordagens (DANTAS, 2019).

A IA emprega várias técnicas, incluindo variantes de busca e otimização matemática, redes neurais artificiais e metodologias baseadas em estatística, probabilidade e economia. É

fundada em várias disciplinas, incluindo ciência da computação, engenharia da informação, matemática, psicologia, idiomas e filosofia (NEVES, 2020).

A área foi criada com a premissa de que o intelecto humano pode ser definido com precisão suficiente para ser imitado por um computador. Isso levanta questões filosóficas sobre a natureza da mente e a ética da construção de entidades artificiais com intelecto semelhante ao humano, tema investigado em mitos, literaturas e filosofia desde a antiguidade (BORGES, 2021).

Alguns indivíduos, como os neoluditas, veem a IA como uma ameaça à humanidade se seu desenvolvimento continuar sem controle. Outros temem que a inteligência artificial, ao contrário das revoluções tecnológicas passadas, represente um perigo de desemprego generalizado (MADEIRA *et al.*, 2020).

As técnicas de IA experimentaram um renascimento no século 21 devido aos avanços no poder do computador, grandes quantidades de dados e compreensão teórica; e as técnicas de IA tornaram-se parte integrante da indústria de tecnologia, auxiliando na resolução de vários problemas difíceis em ciência da computação, engenharia de *software* e pesquisa operacional (DA COSTA *et al.*, 2022).

A definição contemporânea de IA é "o estudo e criação de agentes inteligentes", em que um agente inteligente é um sistema que detecta seu ambiente e otimiza suas chances de sucesso. McCarthy criou a palavra em 1956 e a descreve como a ciência e a engenharia da criação de máquinas inteligentes (PORCHER, 2019).

Outras nomenclaturas propostas para o campo incluem inteligência computacional e racionalidade computacional. A frase inteligência artificial denota a inteligência mostrada por robôs ou programas de computador. Ciência da computação, psicologia, filosofia, neurologia, ciência cognitiva, linguística, pesquisa operacional, economia, teoria de controle, probabilidade, otimização e lógica são todas usadas na pesquisa de IA (ALVES, 2021).

A pesquisa de IA se intersecta com muitas outras áreas, incluindo robótica, sistemas de controle, programação, mineração de dados, logística, reconhecimento de voz e reconhecimento facial, entre outras (MATIAS, 2020).

2.3.1 Raciocínio da IA

Raciocinar é fazer inferências apropriadas à situação. As inferências são classificadas como dedutivas ou indutivas. Um exemplo do primeiro é: “Fred deve estar no museu ou no café. Ele não está no café; portanto, ele está no museu” e, deste último, “acidentes anteriores

desse tipo foram causados por falha do instrumento; portanto, este acidente foi causado por falha do instrumento" (DA COSTA *et al.*, 2022).

A diferença mais significativa entre essas formas de raciocínio é que, no caso dedutivo, a verdade das premissas garante a verdade da conclusão, enquanto no caso indutivo, a verdade da premissa dá suporte à conclusão, mas sem dar garantia absoluta. O raciocínio indutivo é comum na ciência, onde os dados são coletados e modelos experimentais são desenvolvidos para descrever e prever o comportamento futuro - até que o aparecimento de dados anômalos obrigue o modelo a ser revisto (PORCHER, 2019).

Os raciocínios dedutivo e indutivo são comuns na matemática e estatística, onde estruturas elaboradas de teoremas irrefutáveis são construídas a partir de um pequeno conjunto de axiomas e regras básicos. Houve um sucesso considerável na programação de computadores para extrair inferências. No entanto, o verdadeiro raciocínio envolve mais do que apenas desenhar inferências; envolve desenhar inferências relevantes para a solução da tarefa ou da situação específica. Este é um dos problemas mais difíceis que a IA enfrenta (ALVES, 2021).

2.4 APRENDIZADO DE MÁQUINA (*MACHINE LEARNING*)

De acordo com Samuel (1959), "aprendizado de máquina é o campo de estudo que permite que computadores aprendam sem que sejam programados explicitamente". Neste sentido, o termo "aprendizado de máquina" refere-se a "um processo de descobrir um modelo de entrada que faça previsões precisas a partir de dados" (MITCHELL, 1997).

Algumas das inúmeras aplicações possíveis incluem procedimentos diagnósticos automatizados, detecção de fraude de cartão de crédito, análise do mercado de ações, classificação da sequência nucleotídica, reconhecimento de fala e texto e sistemas autônomos.

O aprendizado de máquina tem uma gama ampla de aplicações possíveis. No ambiente da Internet, o aprendizado de máquina é usado, por exemplo, para as seguintes funções (MAYER-SCHÖNBERGER, 2013):

- Detecção independente de e-mails de spam e desenvolvimento de filtros adequados de spam;

- Reconhecimento de voz e texto para assistentes digitais;

- Determinação da relevância de sites para termos de pesquisa;

- Detecção e diferenciação da atividade na Internet entre pessoas físicas e *bots*.

Outras áreas de aplicação para aprendizado de máquina incluem reconhecimento de imagem e rosto, serviços de recomendação automática e detecção automática de fraude de cartão de crédito (DAMACENO *et al.*, 2018).

Há paralelos entre "descoberta de conhecimento em bancos de dados" e "mineração de dados", ambos focados em revelar regularidades e padrões anteriormente despercebidos. Existem vários algoritmos que podem atender a ambos os requisitos. Usando abordagens de "descoberta de conhecimento em bancos de dados", os dados para aprendizado de máquina podem ser criados ou pré-processados (PEDREGOSA, 2011).

Algoritmos são usados para alcançar a implementação prática. Alguns algoritmos de aprendizado de máquina se enquadram em uma das três categorias: aprendizado supervisionado, não supervisionado ou por reforço. Para resumir, os sistemas de tecnologia da informação (TI) podem ser capazes de resolver problemas por conta própria, encontrando padrões em conjuntos de dados existentes (MAYER-SCHNBERGER, 2013).

A área da inteligência artificial que envolve aprendizado de máquina ajuda os sistemas de gestão a desenvolver novas soluções, analisando dados e utilizando algoritmos pré-existentes. O conhecimento baseado em experiência é quase surpreendente em sua natureza sintética, permitindo que novos desafios sejam resolvidos com precisão e dados desconhecidos sejam examinados da mesma forma (ALPAYDIN, 2021).

No entanto, é necessário o envolvimento humano para que o software possa aprender e encontrar soluções sozinho. Antes de iniciar o aprendizado, é preciso fornecer ao sistema dados relevantes e protocolos de treinamento (PEDREGOSA *et al.*, 2011).

Além disso, diretrizes para reconhecimento de padrões e análise de banco de dados devem ser estabelecidas. Com os dados certos e as regras definidas, os sistemas de aprendizado de máquina são capazes de identificar, filtrar e sintetizar informações relevantes, criar hipóteses baseadas em fatos, estimar a probabilidade de ocorrência de um evento e ajustar-se a novas situações por conta própria, ajustando seus processos de acordo com os padrões (MAYER-SCHNBERGER, 2013).

Algoritmos são frequentemente usados em aprendizado de máquina. Eles podem ser classificados em vários tipos, dependendo de como aprendem a ver padrões e resolver problemas (MONARD e BARANAUSKAS, 2003):

- Aprendizagem supervisionada;
- Aprendizado não supervisionado;
- Aprendizagem semi-supervisionada;
- Aprendizado por reforço;
- Aprendizado ativo.

Enquanto no aprendizado supervisionado os modelos de exemplo precisam ser definidos e especificados antecipadamente para combinar as informações com os grupos de modelos dos algoritmos, no aprendizado não supervisionado os grupos de modelos são formados automaticamente com base em padrões reconhecidos independentemente (DAMACENO *et al.*, 2018).

Os grupos de modelos no aprendizado não supervisionado são formados automaticamente com base em padrões reconhecidos independentemente, enquanto os modelos de exemplo no aprendizado supervisionado devem ser definidos e especificados com antecedência para combinar as informações com os grupos de modelos dos algoritmos (BATISTA *et al.*, 2003).

A aprendizagem semi-supervisionada é uma combinação dos tipos de aprendizado supervisionado e não supervisionado, onde alguns dados de treinamento são rotulados e outros são não rotulados. Isso permite que o algoritmo aprenda com uma combinação de informações supervisionadas e não supervisionadas, ampliando a eficiência do aprendizado com menos dados rotulados (CHAPELLE et al., 2006).

Já o aprendizado por reforço envolve o algoritmo aprendendo por meio de recompensas e punições em uma série de ações. O algoritmo toma ações em um ambiente com o objetivo de maximizar a recompensa, ajustando sua estratégia com base nas respostas do ambiente (SUTTON et al., 2018). É uma abordagem amplamente utilizada em problemas de tomada de decisão e inteligência artificial.

Finalmente, o aprendizado ativo oferece ao algoritmo a oportunidade de solicitar os resultados desejados para determinados dados de entrada. Para minimizar o número de

questões, o próprio algoritmo seleciona primeiro questões relevantes com alta relevância para o resultado (DAMACENO *et al.*, 2018).

2.4.1 Tipos de aprendizado

2.4.1.1 Aprendizado Supervisionado

O algoritmo determina uma função para aprender com base em um conjunto de pares de entrada-saída. Para aprender, um "professor" deve fornecer o valor correto para uma função dada uma determinada entrada. O objetivo do aprendizado supervisionado é educar uma rede para vincular dados anteriormente não relacionados, instruindo-a a fazer uma série de cálculos com diferentes entradas e saídas. Um subconjunto do aprendizado supervisionado é o estudo da categorização automática de dados. Um uso prático é o reconhecimento da caligrafia (BATISTA *et al.*, 2003).

Podem ser identificadas algumas subcategorias para aprendizagem supervisionada que são mencionadas com mais frequência na literatura: aprendizagem ativa e autoaprendizagem (MAYER-SCHÖNBERGER, 2013).

2.4.1.2 Aprendizado não Supervisionado

Para um determinado conjunto de entradas, o algoritmo gera um modelo estatístico que descreve as entradas e contém categorias e relacionamentos reconhecidos, permitindo previsões. Existem métodos de agrupamento que dividem os dados em várias categorias que diferem umas das outras em padrões característicos. A rede cria assim independentemente classificadores de acordo com os quais divide o padrão de entrada (BORDA, 2011).

Um algoritmo importante neste contexto é o algoritmo Expectativa-Maximização (EM) que define iterativamente os parâmetros de um modelo de forma a explicar de maneira otimizada os dados vistos. Ele assume a existência de categorias não observáveis e estima alternadamente a pertinência dos dados em uma das categorias e os parâmetros que compõem as categorias. Uma aplicação do algoritmo EM pode ser encontrada, por exemplo, nos Modelos Ocultos de Markov (CAVALLINI, 2008).

Todavia, conforme Mitchell (1997), “o aprendizado não supervisionado é uma forma de aprendizado de máquina na qual o algoritmo é fornecido com uma tarefa sem uma resposta específica correta, e é livre para explorar padrões e relações presentes nos dados."

Além disso, é feita uma distinção entre aprendizado em lote, no qual todos os pares de entrada/saída estão presentes ao mesmo tempo, e aprendizado contínuo (sequencial), no qual a estrutura da rede se desenvolve em momentos diferentes (BORDA, 2011).

Outra diferença é feita entre o aprendizado *on-line*, em que os dados são perdidos após a execução e os pesos foram alterados uma vez, e o aprendizado *off-line*, em que todos os dados são mantidos e podem ser recuperados novamente. Nas abordagens de treinamento, o *batching* é normalmente realizado em segundo plano, mas o treinamento online está em constante progresso (SMOLA *et al.*, 2008).

2.4.1.3 Aprendizado por Reforço

O objetivo do aprendizado por reforço, um ramo do aprendizado de máquina, é como os agentes devem se comportar em um ambiente para maximizar um determinado valor cumulativo de recompensa. Este é um assunto vasto que atrai estudiosos de uma variedade de disciplinas, incluindo, mas não limitado a teoria dos jogos, teoria do controle, pesquisa operacional, teoria da informação, otimização baseada em simulação, sistemas multiagentes, inteligência de enxame, estatística e genética. algoritmos. No aprendizado de máquina, o ambiente geralmente é representado como um processo de decisão de Markov (*Markov Decision Process* - MDP) (DOMINGOS, 2012).

Vários métodos de aprendizado por reforço usam técnicas de programação dinâmica. Quando modelos matemáticos precisos do MDP não são possíveis, técnicas de aprendizado por reforço são aplicadas. Algoritmos de aprendizado por reforço são empregados em carros autônomos e em jogos de aprendizado versus humanos (MONARD e BARANAUSKAS, 2003).

2.4.1.4 Aprendizado Ativo

O aprendizado ativo é uma abordagem de aprendizado de máquina que permite aos modelos selecionarem suas próprias instruções de treinamento para aprimorar seus resultados (SETTLES, 2009). Em vez de depender apenas de dados rotulados fornecidos por um humano, o modelo pode fazer perguntas, coletar dados adicionais e adaptar seu treinamento baseado em sua compreensão atual do problema (SINGH *et al.*, 2010). Isso é útil em situações em que os dados são escassos ou caros de coletar, ou em que as condições do problema mudam ao longo do tempo (COHN *et al.*, 1994).

O aprendizado ativo pode ser usado em combinação com outras técnicas de aprendizado de máquina, como aprendizado supervisionado, não-supervisionado e por reforço (JORDAN, 1992). Em geral, o objetivo do aprendizado ativo é produzir modelos mais eficientes e precisos, que possam realizar tarefas com menos dados e melhor se adaptar a mudanças no ambiente (LEWIS, 1994).

2.4.2 Dados desbalanceados e eventos raros

Os dados desbalanceados são comuns e um problema frequente na análise estatística e na inteligência artificial. Como observaram Chawla *et al.* (2004), "uma distribuição desbalanceada de dados ocorre quando a quantidade de dados em um conjunto é desproporcionalmente maior para algumas classes em relação a outras". Isso afeta a eficiência de alguns algoritmos de análise de dados, já que eles tendem a se ajustar aos dados majoritários.

Outra questão mencionada por Chawla *et al.* (2004) é que "os métodos tradicionais podem não fornecer modelos precisos para dados desbalanceados". Isso deve-se ao fato de que a maioria dos modelos são desenvolvidos para adaptar-se a distribuições uniformes e contínuas. Se os dados não estiverem de acordo com os padrões apropriados, os resultados serão questionáveis e não confiáveis.

Uma possível estratégia para lidar com dados desbalanceados, de acordo com Sun *et al.* (2002), é garantir a existência de diferentes métodos para extrair informações significativas dos dados, podendo ajudar a tornar os resultados mais confiáveis, aumentando a taxa de acerto e diminuindo as chances de cometer erros graves.

Além disso, mencionado por Huang *et al.* (2006), é possível usar técnicas de amostragem para diminuir os efeitos de dados desbalanceados. De modo geral, as técnicas de amostragem são usadas para obter distribuições mais justas e uniformes nos conjuntos de dados, o que ajuda a identificar padrões significativos. Portanto, a amostragem e outras técnicas de estratificação são importantes para o processo de análise de dados e aprendizado de máquina.

Segundo Elkan (2001), "os eventos raros são desafios especiais para o aprendizado de máquina, pois eles são muito menos frequentes do que os eventos comuns." Neste sentido, a detecção de eventos raros é um subconjunto da classificação desequilibrada, onde a classe minoritária é ainda mais rara.

Assim, para lidar com esse problema, diversas técnicas podem ser empregadas, entre elas, a sobreamostragem e a subamostragem (Chawla *et al.*, 2002).

2.4.2.1 Métodos para detecção de eventos raros

De acordo com He e Garcia (2009), uma das abordagens para lidar com problemas de classificação com classes desbalanceadas é a utilização de técnicas de amostragem. Uma dessas técnicas é a sobreamostragem aleatória, que consiste em fazer cópias aleatórias da classe com poucos casos, de forma que as classes fiquem balanceadas. No entanto, essa técnica pode levar a um modelo com *overfitting*, ou seja, o modelo funciona bem com os dados amostrados, mas não garante uma boa taxa de acerto com dados reais. A outra técnica é a subamostragem aleatória, que consiste em excluir observações aleatórias da classe majoritária até que as classes fiquem balanceadas. Essa técnica também pode eliminar observações que seriam importantes para que o modelo aprenda sobre os padrões dos dados.

Para diminuir esses problemas, He e Garcia (2009) sugerem a utilização de técnicas como *Random Over-Sampling Examples* (ROSE) na sobreamostragem aleatória, que consiste em gerar dados sintéticos ao fazer a interpolação dos dados da classe minoritária, e técnicas para manter observações que estejam em regiões fronteiriças com a classe estudada na subamostragem aleatória, a fim de melhorar a representação da região de borda.

3 METODOLOGIA

Neste capítulo, será apresentada a metodologia utilizada para a elaboração deste trabalho de conclusão de curso. Inicialmente, será descrito o percurso metodológico adotado, incluindo a seleção e o tratamento dos dados. Na seção 3.2, serão detalhados os materiais, métodos utilizados e o procedimento de tratamento dos dados. Por fim, serão descritas as métricas de avaliação de desempenho dos modelos aplicados.

O presente trabalho apresenta uma metodologia quantitativa de natureza aplicada e objetivo exploratório. A pesquisa de objetivo exploratório é definida como aquela que proporciona ao pesquisador uma maior familiaridade com o assunto estudado e contribui para construir novas ideias e formar hipóteses (GIL, 2021, p. 26).

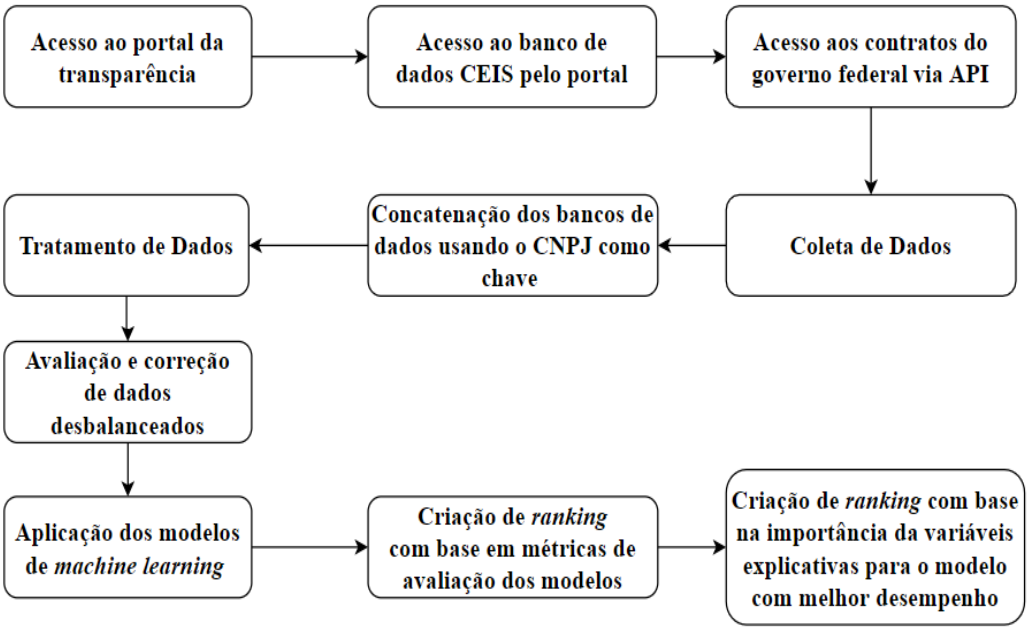
3.1 PERCURSO METODOLÓGICO

O percurso metodológico para a realização dessa pesquisa consistiu, primeiramente, em uma busca para a escolha da base de dados que seria utilizada. Após a escolha e coleta dos dados, foi realizada a concatenação dos bancos de dados, tendo como chave em comum o número do Cadastro Nacional de Pessoas Jurídicas (CNPJ).

Posteriormente, foi realizado o tratamento do banco de dados - procedimento que será abordado no item 3.2.2. Para a realização da correção dos dados desbalanceados, modelagem e obtenção dos resultados, a base de dados foi importada para o software RStudio (RSTUDIO TEAM, 2022), onde utilizou-se a linguagem de programação estatística R (R CORE TEAM, 2022).

Em seguida, procedeu-se à etapa da interpretação dos resultados e, por fim, o penúltimo e último passo do percurso metodológico consistiram na criação de dois *rankings*, um com base nas métricas de avaliação dos modelos de aprendizado de máquina e outro a partir das variáveis explicativas mais importantes para o modelo que obteve melhor desempenho. Todos os passos descritos anteriormente estão ilustrados abaixo na Figura 1.

Figura 1 – Percurso Metodológico



Fonte: Elaboração própria, 2023

3.2 MATERIAIS

3.2.1 Coleta de dados

Os dados utilizados neste estudo foram coletados a partir de duas bases de dados, disponibilizadas por meio do portal API (*Application Programming Interface*) Compras Governamentais e o Portal da Transparência, mantidos, respectivamente, pelo Ministério da Economia (ME) e pela Controladoria Geral da União (CGU).

Os dados do portal API Compras Governamentais referem-se a contratos firmados pela Administração Pública Federal a partir de licitações, enquanto os dados do Portal da Transparência, dizem respeito a empresas cadastradas como inidôneas e/ou suspensas.

Os dados do ME estão disponíveis publicamente por meio de uma API. Devido à limitação de 500 registros por consulta, foi desenvolvido um programa para obter a base completa de contratos licitados, salvando-os em arquivo no formato "*.RDS*" com 642.000 observações e 21 atributos. Cada observação representa um contrato diferente. A extração dos dados foi executada no dia 07/01/2022 e está detalhada no Anexo A.

Ressalta-se que um arquivo no formato "*.RDS*" consiste num banco de dados serializado, ou seja, é um arquivo que contém uma representação binária de um objeto R (R

CORE TEAM, 2022), geralmente uma *data frame* ou uma matriz. Ele é uma opção comum para salvar e carregar dados estruturados no R de forma eficiente.

O conjunto de informações do Portal da Transparência aborda empresas e pessoas físicas que sofreram sanções tendo como efeito a restrição ao direito de participar de licitações ou de celebrar contratos com a Administração Pública. Essa base de dados é chamada de Cadastro Nacional de Empresas Inidôneas e Suspensas (CEIS) e apresentou 15.522 observações e 22 atributos. Ela está disponível para *download* no formato *Comma-separated values* (CSV) no Portal da Transparência. O *download* do banco de dados foi feito no dia 07/01/2022 por meio do endereço <https://portaldatransparencia.gov.br/download-de-dados/ceis >.

3.2.2 Tratamento dos dados

Inicialmente, foi realizada a concatenação da base de contratos e a base CEIS, tendo como elemento-chave, o número do CNPJ. Prosseguiu-se com a limpeza, validação e organização dos dados coletados. O procedimento consistiu na remoção de dados faltantes, seleção das variáveis mais importantes, conforme avaliação a critério do pesquisador, garantindo que os dados estivessem estruturados de forma consistente. O detalhamento do *script* do procedimento encontra-se no Apêndice C.

Após a limpeza e seleção das variáveis, resultando em um repositório de 610.840 observações e 23 variáveis, detalhadas na Tabela 1:

Quadro 1 – Variáveis antes do balanceamento dos dados

Variáveis
Identificador do Contrato
UASG
Modalidade da Licitação
Origem da Licitação
Número
Objeto
Número de Aditivos
CPF Contratada
CNPJ Contratada
Data de Assinatura
Fundamento Legal
Data de Início da Vigência

Data de Termino da Vigência
Valor
cnpj
nome
tipo_de_pessoa
raz_o_social_cadastro_receita
tipo_sanção
abragência_definida_em_decisão_judicial
fundamentação_legal
EIS
uasg2

Fonte: Elaboração própria, 2023

3.2.2.1 Avaliação e correção de dados desbalanceados

Com base na Revisão Bibliográfica, foi possível compreender a variável resposta como um evento raro. O grande desequilíbrio no número de observações de empresas idôneas e inidôneas ou suspensas tornou imperativo o balanceamento dos dados, tendo em vista a tabela 1.

Tabela 1 – Número de instâncias antes do balanceamento dos dados

Empresas idôneas	Empresas inidôneas ou suspensas
585.024	25.816

Fonte: elaborada pelo autor, 2023.

Para tal fim, foi utilizado o pacote ROSE, gerando dados sintéticos por meio da interpolação dos dados da classe minoritária. Deste modo, corrigiu-se o desbalanceamento, resultando em uma base de dados equilibrada com 1.163.418 observações.

Tabela 2 – Número de instâncias após balanceamento dos dados

Empresas idôneas	Empresas inidôneas ou suspensas
581398	581750

Fonte: elaborada pelo autor, 2023.

3.2.3 Descrição das variáveis

O banco de dados final apresentou 5 variáveis detalhadas a seguir:

Variável de interesse/ Variável Resposta

- Empresa inidônea ou suspensa (EIS): variável resposta categórica com duas possíveis classificações, "inidônea ou suspensa", onde atribuiu-se o valor 1 ou "não inidônea", com valor 0, transformando-a em variável *dummy*.

Preditores lineares

- Modalidade: variável preditora categórica, representa uma categoria de modalidade de licitação, sendo categorizados como “Convite”, “Tomada de Preços”, “Concurso”, “Tomada de Preços por Técnica e Preço”, “Concorrência”, “Concorrência por Técnica e Preço”, “Concorrência Internacional”, “Pregão”, “Dispensa de Licitação”, “Inexigibilidade De Licitação” e “RDC”.

- Aditivos: variável preditora numérica, representa um valor quantitativo do número de aditivos feitos em um contrato.

- Valor dos contratos: variável preditora numérica, representa o valor monetário do contrato.

- UASG: variável preditora categórica, representa uma macro área do Governo Federal, após agrupamento de classificações da base de dados inicial, sendo categorizadas como “Administração”, “Agricultura”, “Educação”, “Militar”, “Procuradoria” e “Outros”.

3.2.4 Modelagem

Nesta etapa, é realizada uma separação aleatória de 80% dos dados, totalizando 930.518 observações para compor o conjunto de treinamento e outras 232.630 observações são reservadas para testar os modelos.

Esta técnica é conhecida como validação cruzada *holdout* e tem como objetivo separar uma amostra dos dados para que ela não seja considerada durante a etapa de treinamento do modelo. Essa amostra é posteriormente utilizada para verificar se o resultado obtido durante o aprendizado é semelhante ao esperado em situações reais.

Foram aplicadas seis técnicas diferentes de aprendizado de máquina: Regressão Logística, *Random Forest*, Redes Neurais, *Naïve Bayes* e *Stochastic Gradient Boosting* e Árvore de Decisão. Essas técnicas serão explicadas na seção subsequente. Foi usada a linguagem R (R CORE TEAM, 2022) e o pacote *caret* (KUHN, 2008*)*, utilizado comumente para treinar modelos de classificação e regressão, pois fornece uma interface unificada para vários algoritmos de aprendizado de máquina.

3.3 MÉTODOS

Nesta seção, serão descritos os modelos de *aprendizado de máquina* empregados para resolver o problema proposto. As técnicas utilizadas incluem regressão logística, *Random Forest*, redes neurais, *Naïve Bayes*, *Stochastic Gradient Boosting* e árvore de decisão. Cada uma dessas técnicas será descrita em detalhe. A escolha dessas técnicas foi baseada em sua capacidade de lidar com problemas de classificação binária e multi-classificação, bem como sua capacidade de lidar com grandes conjuntos de dados.

3.3.1 Árvore de decisão

A árvore de decisão é uma técnica de aprendizado supervisionado que pode ser usada tanto para problemas de classificação quanto de regressão, mas é preferida principalmente para resolver problemas de classificação. Esta técnica consiste em uma estrutura de árvore, onde os nós internos representam as características de um conjunto de dados, os ramos representam as regras de decisão e cada nó folha representa o resultado. Em uma árvore de decisão, existem dois nós, que são o nó de decisão e o nó folha (BUNTINE, 2020).

Os nós de decisão são responsáveis por tomar decisões e possuem várias ramificações, enquanto os nós folha são as saídas destas decisões, não possuindo ramificações adicionais. As decisões ou os testes são realizados com base nas características do conjunto de dados fornecido, criando uma representação gráfica de todas as soluções possíveis para um problema ou decisão (ABDELHALIM e TRAORE, 2009).

Ademais, novas árvores de decisão são adicionadas ao modelo para corrigir o erro residual do modelo existente. Cada árvore de decisão é criada usando o algoritmo guloso (*greed)* para selecionar os pontos de divisão que melhor minimizam uma função objetivo. Isso pode resultar em árvores que usam os mesmos atributos e até mesmo os mesmos pontos de divisão repetidas vezes (ABDELHALIM e TRAORE, 2009).

A árvore de decisão imita a capacidade humana de tomar decisões, sendo de fácil compreensão devido à sua estrutura semelhante à de uma árvore. Para prever a classe do conjunto de dados, o algoritmo começa no nó raiz da árvore, comparando os valores do atributo da raiz com o atributo do conjunto de dados real e seguindo o ramo correspondente (BUNTINE, 2020). Este processo é repetido até que o nó folha seja atingido.

A seleção do melhor atributo para os nós da árvore é uma questão relevante na implementação da árvore de decisão. Para resolver este problema, existe a técnica de medida de seleção de atributos (ASM), que permite selecionar o atributo mais adequado para os nós. Duas técnicas populares de ASM são o ganho de informação e o índice de Gini (ABDELHALIM e TRAORE, 2009).

O ganho de informação é a medida das mudanças na entropia após a segmentação de um conjunto de dados com base em um atributo. Ele calcula quanta informação um recurso nos fornece sobre uma classe. O índice de Gini é uma medida de impureza usada durante a criação de uma árvore de decisão no algoritmo CART (*Classification and Regression Tree*). Um atributo com baixo índice de Gini deve ser preferido em comparação com o alto índice de Gini (BUNTINE, 2020).

3.3.2 *Random Forest*

A floresta aleatória *(Random Forest)* é um algoritmo de aprendizado de máquina Supervisionado amplamente utilizado em problemas de Classificação e Regressão. Ele constrói florestas de decisão em diferentes amostras e leva sua maioria de votos para classificação e média em caso de regressão. Uma das características mais importantes da *random forest* é que ela pode lidar com um conjunto de dados contendo variáveis contínuas, como no caso de regressão e variáveis categóricas como no caso de classificação (CUTLER, CUTLER e STEVENS, 2012).

A *random forest* gera resultados treinados por meio do algoritmo *bagging*, também conhecido como *bootstrap aggregation*. Este método escolhe uma amostra aleatória dos dados e cada modelo é gerado a partir dessas amostras com substituição, chamadas de amostras

bootstrap (BREIMAN, 2001). A saída final é baseada na votação por maioria após a combinação dos resultados de todos os modelos (HASTIE, TIBSHIRANI E FRIEDMAN, 2009).

Ademais, o maior número de árvores no algoritmo de *random forest* leva a uma maior precisão e evita o problema de *overfitting*. Como o *random forest* combina várias árvores para prever a classe do conjunto de dados, é possível que algumas árvores de decisão possam prever a saída correta, enquanto outras não, mas juntas, todas as árvores apresentam uma saída com melhor resultado (BREIMAN, 2001).

A *Random Forest* funciona em duas fases: a primeira é criar a floresta aleatória combinando *N* árvores de decisão e a segunda é fazer previsões para cada árvore criada na primeira fase. Ademais, existem principalmente quatro setores onde a *Random Forest* é mais usada (CUTLER, CUTLER e STEVENS, 2012).

Primeiramente, o setor bancário usa principalmente esse algoritmo para a identificação do risco de empréstimos. Como também, com a ajuda deste algoritmo, as tendências e os riscos da doença podem ser identificados pela medicina. Pode se identificar as áreas de uso da terra semelhante e as tendências de marketing (HASTIE, TIBSHIRANI E FRIEDMAN, 2009).

3.3.3 *Stochastic Gradient Boosting*

Algoritmo de aprendizado de máquina que utiliza uma técnica de *boosting* para ajustar sequencialmente modelos fracos (geralmente árvores de decisão) para melhorar a precisão do modelo final.

O *Stochastic Gradient Boosting* é uma técnica de aprendizado de máquina baseada na ideia de que a combinação de modelos anteriores com o melhor modelo possível minimiza o erro geral de previsão. A chave para esse processo é estabelecer resultados alvo para o próximo modelo, visando minimizar o erro (NASSIF, 2016).

3.3.4 *Naïve Bayes*

O *Naïve Bayes* é um algoritmo de aprendizado supervisionado, baseado no teorema de Bayes e usado para resolver problemas de classificação. É amplamente utilizado para resolver problemas de classificação de texto em conjuntos de dados de alta dimensão. É conhecido como

um dos algoritmos de classificação mais simples e eficazes, proporcionando modelos rápidos de aprendizado de máquina capazes de fazer previsões rápidas (XU, 2018).

Recebe o nome de *Naïve* devido à sua suposição de independência entre as características e Bayes por se basear no princípio do Teorema de Bayes (LEUNG, 2007), também conhecido como Regra de *Bayes* ou Lei de *Bayes*, este teorema é usado para determinar a probabilidade de uma hipótese com conhecimento prévio, dependendo da probabilidade condicional (WEBB, KEOGH e MIIKKULAINEN, 2010). A fórmula do teorema de *Bayes* é dada por:

$$P(A/B) = \frac{P(B/A)xP(A)}{P(B)} \qquad \text{Eq.1}$$

Existem três tipos de modelos de *Naïve Bayes*. O modelo gaussiano assume que os recursos seguem uma distribuição normal. Isso significa que, se os preditores assumirem valores contínuos em vez de discretos, o modelo assumirá que esses valores são amostrados da distribuição gaussiana (WEBB, KEOGH e MIIKKULAINEN, 2010).

O classificador multinominal *Naïve Bayes* é usado quando os dados são distribuídos multinominal. É usado principalmente para problemas de classificação de documentos, indicando que um determinado documento pertence a qual categoria, como esportes, política, educação etc. O classificador usa a frequência de palavras para os preditores (RISH *et al.*, 2001).

O classificador de Bernoulli funciona de maneira semelhante ao classificador multinominal, mas as variáveis preditoras são as variáveis booleanas independentes. Como se uma determinada palavra estivesse presente ou não em um documento. Este modelo também é famoso por tarefas de classificação de documentos (LEUNG, 2007).

O *Naïve Bayes* possui vantagens como ser rápido e fácil para prever classes de dados, ser adequado para classificações binárias e multiclasse, ter bom desempenho em previsões de várias classes, ser aplicável na pontuação de crédito e classificação de dados médicos, permitir previsões em tempo real e ser usado na classificação de texto, como filtragem de spam e análise de sentimento.

3.3.5 Redes Neurais

As redes neurais, também conhecidas como redes neurais artificiais (RNAs), formam um subconjunto do aprendizado de máquina e estão no centro dos algoritmos de aprendizado profundo. Seu nome e estrutura são inspirados no cérebro humano, imitando a maneira como os neurônios biológicos sinalizam uns aos outros. Camadas de nós compõem redes neurais artificiais, incluindo uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída (FLECK *et al.*, 2016).

Cada nó, ou neurônio artificial, está conectado a outro e tem um peso e um limiar associados a ele. Se a saída de qualquer nó individual estiver acima do valor limite especificado, esse nó é ativado, enviando dados para a próxima camada da rede. Caso contrário, nenhum dado é repassado para a próxima camada da rede (DE CASTRO e DE CASTRO, 2001).

As redes neurais dependem de dados de treinamento para aprender e melhorar sua precisão ao longo do tempo (RAUBER, 2005). No entanto, uma vez que esses algoritmos de aprendizado são ajustados para precisão, eles são ferramentas poderosas em ciência da computação e inteligência artificial, permitindo classificar e agrupar dados em alta velocidade. Tarefas de reconhecimento de fala ou reconhecimento de imagem podem levar minutos ou horas quando comparadas à identificação manual por especialistas humanos (FLECK *et al.*, 2016).

Em situações da vida real, as redes neurais podem ajudar as pessoas a resolver problemas complexos. Como resultado, eles podem aprender e modelar as relações entre entradas e saídas que são não lineares e complexas. Eles também podem fazer generalizações e inferências e descobrir relacionamentos, padrões e previsões ocultos (KOVÁCS, 2002).

As redes neurais também modelam dados altamente voláteis (como dados de séries temporais financeiras) e variações necessárias para prever eventos raros (como detecção de fraude). As redes neurais podem melhorar a tomada de decisões em áreas como: diagnóstico médico e de doenças, otimização da logística para redes de transporte, carga elétrica e previsão de demanda de energia e previsões financeiras para preços de ações, moedas, opções, futuros, falências e classificações de títulos (RAUBER, 2005).

Na programação convencional, os dados são armazenados na rede em vez de um banco de dados. As redes neurais garantem que toda a operação da rede não seja interrompida quando alguns dados desaparecem de um local (DE CASTRO e DE CASTRO, 2001). Como também, fornecem boa tolerância a falhas. Eles garantem que a corrupção de uma ou várias células de

rede artificial não afete a produção de saída. Portanto, as redes podem tolerar melhor os erros (FLECK *et al.*, 2016).

3.3.6 Regressão logística

No aprendizado de máquina, a regressão logística é um modelo estatístico usado para estudar as relações entre um conjunto de variáveis quantitativas e qualitativas Xi e uma variável qualitativa Y (COSTALAT e TAVARES, 2022).

De fato, a regressão logística é um modelo linear generalizado que usa uma função logística como função de ligação. Tornou-se ao longo dos anos uma ferramenta estatística indispensável na disciplina de aprendizado de máquina. Além disso, é considerado um dos modelos de análise multivariada mais fáceis de analisar. Pode assumir várias formas, nomeadamente a forma logística ou linear, mas também a forma binária ou multinominal (SCHNEIDER, 2016).

Por outro lado, se esse valor for maior que o mesmo limite inicial, é provável que o evento ocorra. É importante notar que o resultado pode ser interpretado como uma probabilidade que varia sempre entre 0 e 1 (SCHNEIDER, 2016).

3.4 MÉTRICAS DE AVALIAÇÃO DE DESEMPENHO DOS MODELOS DE *APRENDIZADO DE MÁQUINA*

Para avaliar o desempenho de modelos de aprendizado de máquina, utilizam-se diversas métricas, que dependem do tipo de problema sendo tratado. De acordo com Martínez (2012), “a métrica de avaliação de um modelo de aprendizado de máquina deve ser escolhida de acordo com o tipo de problema sendo tratado e o tipo de modelo escolhido para o trabalho”. Serão discutidas as métricas utilizadas para criação de ranking dos modelos em capítulo posterior.

3.4.1 Matriz de Confusão

A matriz de confusão é um instrumento importante para avaliar o desempenho de sistemas de classificação (Kuhn, 2014). Seu uso é comum para aferir a precisão dos modelos gerados para problemas de classificação binária, mas também pode ser aplicado a problemas de múltiplas classes (LAPEDES E FARBER, 1987).

De acordo com a teoria estatística de decisão, a matriz de confusão é uma quantificação dos erros cometidos na classificação (Kohavi et al., 1998). Ela fornece uma síntese de informações, apresentando uma visão geral das classificações realizadas e sua correspondência com as ações reais, e pode ser uma fonte valiosa de informações para os usuários.

A seguir, é apresentado uma possível representação da matriz de confusão:

Quadro 2 – Matriz de Confusão

		Valor predito	
		Sim	Não
Valor Real	Sim	VP	FN
	Não	FP	VN

Fonte: Elaboração própria, 2023

em que VP significa verdadeiro positivo, FP é falso positivo, FN é falso negativo e VN é verdadeiro negativo.

Ela permite não apenas avaliar a precisão, como também verificar a uniformidade dos erros (CORTES *et al.*, 1995), podendo revelar informações valiosas sobre o valor dos modelos de classificação.

Existe um conjunto de métricas que podem ser formuladas para medir o desempenho de classificadores com base na Matriz de Confusão. Entre eles, destaca-se a acurácia, a precisão e a revocação (SOKOLOVA *et al.,* 2009).

3.4.1.1 Acurácia

A acurácia tem sido definida como a capacidade de um sistema, algoritmo ou modelo de produzir resultados precisos. Segundo Revu *et al.* (2017), ela se baseia na comparação entre saídas previstas e resultados reais, e está principalmente relacionada à avaliação de modelos preditivos. Enquanto a acurácia tem se mostrado eficaz para classificação binária, ela não é absoluta. Zarepour *et al.* (2016) acreditam que a acurácia, sozinha, não representa o erro de acerto/falha, o que pode ser limitante em aplicações que necessitem de mais detalhes sobre erros específicos. Por exemplo, se uma saída espera apenas duas classes sim/não, a acurácia pode

expressar campos variados na modelagem com certa facilidade, mas quando necessário diferenciar entre mais variáveis, ela se torna insuficiente.

Nesse caso, ter métricas que levem em consideração erros e acertos específicos se torna importante. De maneira geral, Garem *et al.* (2020) acreditam que é importante entender que a acurácia por si só não pode abranger todos os perfis de erros e acertos, e é preciso contar com outras métricas para fazer esse tipo de análise.

A acurácia pode ser representada pela fórmula:

$$Acurácia = \frac{VP + VN}{(VP + FN + VN + FP)} \quad \text{Eq.2}$$

3.4.1.2 Especificidade

A especificidade é definida como a proporção de verdadeiros negativos (ou seja, o número de casos em que o modelo prevê corretamente que uma determinada classe não está presente) em relação ao número total de negativos (ou seja, o número de casos em que a classe não está presente).

Além disso, de acordo com Geman e Geman (1984), especificidade ainda pode ser utilizada para avaliar outras características importantes, como a capacidade de um modelo de classificação em estabelecer limites aceitáveis para Falsos Positivos.

$$Especificidade = \frac{VN}{(FP + VN)} \quad \text{Eq.3}$$

3.4.1.3 *Recall*

O *recall*, também conhecido como sensibilidade ou revocação, é um método para realizar o ajuste de classificação. De acordo com Bühlmann e van de Geer (2014), "A revocação de um classificador pode servir como uma medida de performance útil, já que nos fornece informações sobre a precisão da classificação, além da recuperação do conjunto inteiro dos dados".

Consiste na seguinte relação:

$$Recall = \frac{VP}{(VP + FN)} \quad \text{Eq.3}$$

3.4.1.4 Precisão

Segundo Tibshirani *et al.* (2003), a precisão é uma medida de desempenho para classificações. Ela consiste em calcular o percentual de pontos de dados previamente conhecidos que o modelo é capaz de classificar corretamente.

De acordo com Landgrebe (2002), a precisão é muitas vezes usada para avaliar modelos de classificação como árvores de decisão, KNN e SVM, pois seu cálculo é direto e não gera restrições de parâmetro que demandem escolhas subjetivas.

Além disso, Hellge *et al.* (1998) afirmam que as pontuações de precisão oferecem uma avaliação geral aproximada boa e rápida dos resultados e destacam-se pela sua facilidade de interpretação.

Pode ser expressado por:

$$Precisão = \frac{VP}{(VP + FP)} \quad \text{Eq.4}$$

em que VP consiste no número de verdadeiros positivos e FP corresponde ao número de falsos positivos.

3.4.1.5 *F1-score*

O *F1-score* é uma métrica utilizada para medir o desempenho de classificadores binários ou que retornem resultados 0 ou 1. Como afirma Souto *et al.* (2016), o *F1-score* é normalmente considerado uma métrica adequada para capturar o desempenho geral dos classificadores. Ele é uma média ponderada entre precisão e *recall*. A fórmula para calcular o *F1-score* é:

$$F1 = \frac{2 * Precisão * Recall}{Precisão + Recall} \quad \text{Eq.5}$$

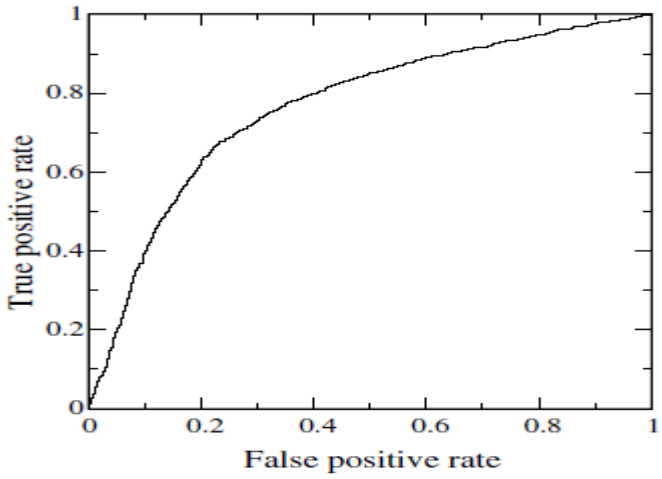
Devido ao fato de que a precisão e o *recall* são valores que variam entre 0 a 1, o resultado do *F1-score* também estará situado dentro desse intervalo, sendo que um resultado próximo de 0 indica um baixo desempenho nos classificadores e um valor próximo de 1 mostra um alto desempenho.

3.4.1.6 Curva ROC e a AUC

A curva ROC fornece informações abrangentes sobre o desempenho do classificador, medindo a correlação entre a distribuição das predições e as classes originais dos dados (ZHANG *et al.*, 2004).

Segundo Flach (2014), a curva ROC é utilizada para situações em que é preciso decidir entre duas possibilidades, isto é, qual é o limiar entre duas classes. Para Fawcett (2006), conforme figura 2, a curva ROC "representa visualmente as avaliações de todos os possíveis limiares de probabilidade de classificação", tendo como medida padrão para avaliar o desempenho das curvas ROC, a AUC (*Area Under ROC Curve*).

Figura 2 – Exemplo de curva ROC



Fonte: Fawcett (2006)

De acordo com Rijsbergen (1979), a AUC descreve o quão bem um classificador se sai em diferentes áreas de aplicação e permite comparações com o desempenho de outros classificadores para o mesmo conjunto de dados. Portanto, espera-se que um bom classificador tenha alta AUC para uma variada seleção de sinais e níveis de limiar.

4 RESULTADOS

Neste capítulo, serão apresentados e analisados os resultados obtidos da aplicação dos algoritmos de aprendizado de máquina previamente selecionados para solução do problema proposto. A análise dos resultados é de extrema importância para avaliar a eficiência dos modelos em questão, bem como para a compreensão da relação entre os dados coletados e o problema.

As métricas utilizadas na avaliação dos modelos foram escolhidas de acordo com as melhores práticas da literatura, permitindo uma comparação objetiva entre os modelos e uma avaliação da qualidade dos resultados. As métricas utilizadas incluem acurácia, especificidade, precisão, *recall* e *F1-score* e AUC.

Ressalta-se que todos os modelos tiveram o desempenho avaliado na mesma base reservada para testes, conforme descrito no item 3.2.4 deste trabalho.

4.1 INDICADORES

Com base nas matrizes de confusão detalhadas no anexo A e na aplicação das métricas expostas e descritas, foi possível observar que a *random forest* se apresentou como o algoritmo de aprendizado de máquina que obteve os melhores resultados em todas as métricas avaliadas.

Em relação à acurácia, calculada em 0,9522, o que significa que ele foi capaz de acertar 95,22% das previsões realizadas e indica uma alta capacidade de generalização do modelo, ou seja, sua habilidade em realizar previsões precisas em dados não vistos durante o treinamento.

Quanto à especificidade, o modelo identificou corretamente 92,02% das instâncias negativas. Este resultado indica que o modelo possui uma boa capacidade de evitar falsos negativos.

A precisão do modelo foi de 0,9253, indicando que das previsões positivas realizadas, 92,53% foram corretas. Este resultado indica a habilidade do modelo em realizar previsões precisas, evitando falsos positivos.

O *recall* do modelo foi de 0,9841, o que indica que ele identificou 98,41% das instâncias positivas. Este resultado indica a excelente habilidade do modelo em identificar todas as instâncias positivas, evitando falsos negativos.

Por fim, o *F1-score* do modelo foi de 0,9538, indicando que o modelo apresentou um desempenho equilibrado em termos de precisão e *recall*.

Em resumo, a *random forest* foi o algoritmo que apresentou os melhores resultados em todas os indicadores avaliados, apontando que se trata um modelo de classificação altamente eficiente e preciso, conforme se observa detalhadamente a seguir:

Quadro 3 – *Ranking* dos modelos de *aprendizado de máquina* ordenados do melhor para o pior desempenho em cada indicador

Indicador	Modelo	Resultado
Acurácia	***Random Forest***	**0.9522**
	Stochastic Gradient Boosting	0.6273
	Regressão Logística	0.613
	Árvore de Decisão	0.6089
	Naïve Bayes	0.6062
	Redes Neurais	0.6036
Especificidade	***Random Forest***	**0.9202**
	Regressão Logística	0.6995
	Stochastic Gradient Boosting	0.4153
	Redes Neurais	0.3531
	Naïve Bayes	0.3386
	Árvore de Decisão	0.3327
Precisão	***Random Forest***	**0.9253**
	Regressão Logística	0.8315
	Stochastic Gradient Boosting	0.5901
	Árvore de Decisão	0.5708
	Naïve Bayes	0.5698
	Redes Neurais	0.5697
Recall	***Random Forest***	**0.9841**
	Árvore de Decisão	0.8841
	Naïve Bayes	0.8728
	Redes Neurais	0.8532
	Stochastic Gradient Boosting	0.8386
	Regressão Logística	0.6062
F1-score	***Random Forest***	**0.9538**
	Árvore de Decisão	0.6937
	Stochastic Gradient Boosting	0.6927
	Naïve Bayes	0.6895
	Redes Neurais	0.6832
	Regressão Logística	0.6828
AUC	***Random Forest***	**0.9624**

	Regressão Logística	0.6383
	Stochastic Gradient Boosting	0.6273
	Árvore de Decisão	0.6082
	Naïve Bayes	0.6052
	Redes Neurais	0.6025

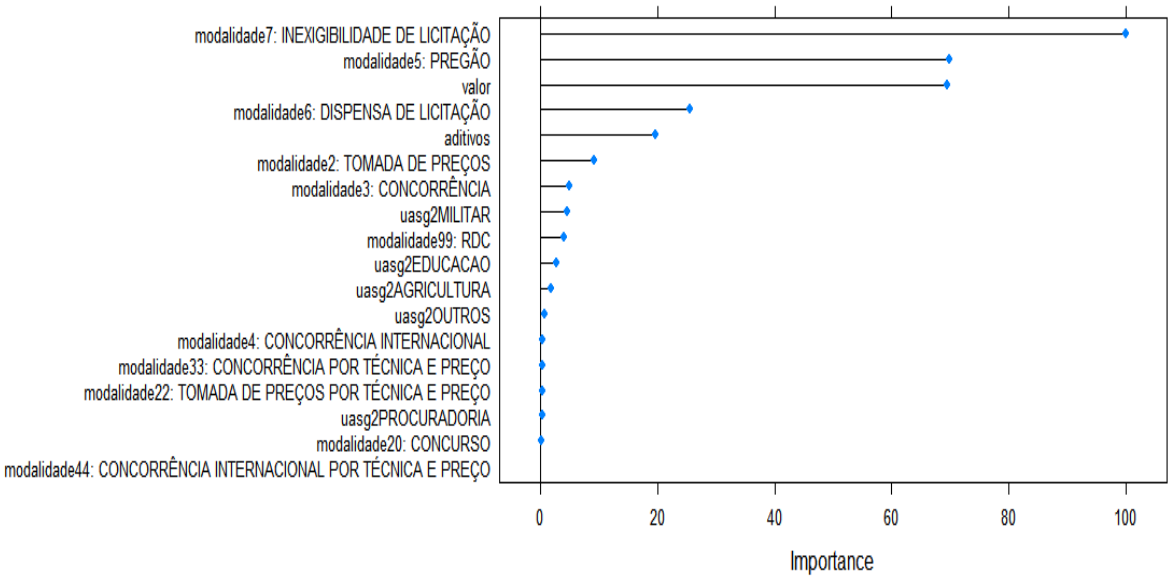
Fonte: Elaboração própria, 2023.

A partir do modelo de melhor desempenho, avaliou-se as variáveis mais importantes para explicar a ocorrência de fraudes. A análise das variáveis importantes para o modelo pode ser feita por um especialista no processo de compras governamentais a fim de criar novas regras para a redução de fraudes.

4.2 VARIÁVEIS MAIS IMPORTANTES PARA IDENTIFICAR EMPRESAS INIDÔNEAS

As variáveis inexigibilidade de licitação, pregão, a modalidade de licitação mais utilizada, e o valor, apresentaram maior importância, conforme gráfico abaixo. Os resultados são de suma relevância, à medida que, conforme exposto, a regra geral é a realização de licitações pelo Poder Público, sendo dispensável em caráter excepcional. Por fim, hipóteses esperadas como a importância do valor e a modalidade pregão foram confirmadas.

Figura 2 – Gráfico de importância das variáveis



Fonte: Elaboração própria, 2023.

5 CONCLUSÃO

Neste trabalho, foi realizada uma avaliação sobre a viabilidade de identificar empresas fraudadoras em licitações públicas por meio de técnicas de aprendizado de máquina, tendo como fonte de informações os dados disponibilizados de contratos firmados pelo governo federal e os cadastros de empresas inidôneas ou suspensas.

Em princípio, uma importante etapa de pré-processamento dos dados foi realizada, visando assegurar a qualidade e a adequação das bases de dados para a construção dos modelos. Esta etapa incluiu uma série de tarefas, tais como a verificação da integridade dos dados, a correção de erros e a remoção de duplicatas para garantir que estivessem em formato consistente.

Este trabalho de pré-processamento dos dados é crucial para garantir a qualidade e a confiabilidade dos resultados dos modelos, bem como para maximizar a eficiência e a eficácia do processo de construção dos modelos. Além disso, a realização desta etapa pôde ajudar a minimizar o risco de viés ou erros nos resultados, garantindo assim a robustez e a validade dos modelos construídos.

Foi possível concluir que modelos de aprendizado de máquina aplicados ao problema de pesquisa são capazes de classificar e realizar previsões com desempenho robusto. Esses modelos permitem que órgãos de governança priorizem investigações em licitações com maior probabilidade de estarem sendo fornecidas por empresas fraudulentas.

A análise dos resultados também revelou a inexigibilidade de licitação, o emprego de pregão e o valor do contrato como fatores importantes para ocorrência de fraudes, o que sugere a possibilidade de mudanças em processos governamentais que favoreçam essas modalidades de compras em detrimento de outros e maior atenção aos valores.

Para futuras pesquisas, sugere-se a normalização de variáveis numéricas a fim de melhorar o desempenho dos modelos, uso de métodos de seleção de variáveis como a Regressão por *Stepwise*, a definição adequada de hiper parâmetros e a incorporação de mudanças na legislação de licitações para melhorar a performance dos modelos de aprendizado de máquina.

Em geral, este trabalho indica que o uso de técnicas de aprendizado de máquina na administração pública pode fornecer importantes avanços na eficiência e efetividade da gestão pública, e é importante que futuras pesquisas explorem essas possibilidades de maneira ainda mais ampla.

REFERÊNCIAS

ABDELHALIM, Amany; TRAORE, Issa. **A new method for learning decision trees from rules.** In: 2009 International Conference on Machine Learning and Applications. IEEE, 2009. p. 693-698.

AGUIAR, F. G. **Utilização de Redes Neurais para Detecção de Padrões de Vazamentos.** Universidade de São Paulo. São Carlos. 2010.

ALPAYDIN, E. **Introduction to Machine Learning**, 2nd edn. Adaptive Computation and Machine Learning. [S.l.]: The MIT Press (February 2010), 2010

ALVES, Marcus Felipe Coelho. **Modelagem de processos para viabilizar a implementação de automação robótica no processo (RPA) de atendimento em um órgão do governo do Estado do Ceará.** UNICHRISTUS, 2021.

BARROSO, Sérgio Luiz; BARROSO Henrique Gabriel. **O que é e como funciona o processo de licitação?.** JusBrasil, 2017 Disponível em <https://sergioluizbarroso.jusbrasil.com.br/artigos/437627975/o-que-e-e-como-funciona-o-processo-de-licitacao >. Acesso em 23 de novembro de 2022.

BATISTA, Gustavo Enrique de Almeida Prado *et al*. **Pré-processamento de dados em aprendizado de máquina supervisionado**. 2003. Tese de Doutorado. Universidade de São Paulo.

BORDA, M. **Statistical and informational model of an its**. In: Fundamentals in Information Theory and Coding. [S.l.]: Springer, 2011.

BORGES, Danihanne. **A influência das ferramentas big data e inteligência artificial no marketing 4.0.** Research, Society and Development, v. 10, n. 5, p. e50210515296-e50210515296, 2021.

BRASIL. **Constituição Da República Federativa Do Brasil De 1988**. Disponível em <https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Acesso em 23 de novembro de 2022.

BRASIL. **Decreto-lei nº 2.848, de 7 de dezembro de 1940**. Disponível em <https://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm>. Acesso em 23 de novembro de 2022.

BRASIL. **Lei nº 10.520, de 17 de julho de 2002**. Disponível em <https://www.planalto.gov.br/ccivil_03/leis/2002/l10520.htm>. Acesso em 23 de novembro de 2022.

BRASIL. **Lei nº 12.529, de 30 de novembro de 2011**. Disponível em <https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12529.htm>. Acesso em 23 de novembro de 2022.

BRASIL. **Lei nº 8.666, de 21 de junho de 1993.** Disponível em <https://www.planalto.gov.br/ccivil_03/leis/l8666cons.htm>. Acesso em 23 de novembro de 2022.

BRASIL. **Súmula de jurisprudência, enunciado nº 331 (Contrato de prestação de serviços. Legalidade).** DEJT 27, 30 e 31.05.2011. Disponível em: <http://www.tst.jus.br/jurisprudencia/Sumulas_com_indice/Sumulas_Ind_301_350.html>. Acesso em 23 de novembro de 2022.

BREIMAN, Leo. **Random Forests.** Machine learning, v. 45, n. 1, p. 5-32, 2001.

BRINK, H.; RICHARDS, J. W.; FETHEROLF, M. **Real-world machine learning**. [S.l.: s.n.], 2015.

BÜHLMANN, Peter; GEER, Sara van de. Statistics for High-Dimensional Data. New York: Springer, 2014.

BUNTINE, Wray. **Learning classification trees**. In: Artificial Intelligence frontiers in statistics. Chapman and Hall/CRC, 2020. p. 182-201.

CAVALLINI, R. **O Marketing depois de amanhã**. São Paulo: Ed. do Autor, 2008.

CAZELLA, Carla. **Fraudes em licitações**: como identificar e ajudar a combate-las. Anuário Pesquisa e Extensão Unoesc Chapecó, v. 4, p. e21055-e21055, 2019.

CHAWLA, Nitesh V.; BOWYER, Kevin W.; HALL, Lawrence O.; KEGELMEYER, W. Philip. **SMOTE: Synthetic Minority Over-sampling Technique.** Journal of Artificial Intelligence Research, v. 16, p. 321-357, 2004.

CHENISK, Diego Ari. **Distinção entre modalidade e tipo de licitação**. Disponível em <http://www.migalhas.com.br/dePeso/16,MI67167,11049-Distincao+entre+modali dade+e+tipo+de+licitacao>. Acesso em 23 de novembro de 2022.

COHN, D. A.; ATLAS, L.; LADNER, R. E. Active learning. Machine learning, v. 15, n. 2, p. 201-256, 1994.

CORTES, C.; VAPNIK, V. Support-vector networks. Machine Learning, n.20, p.273-297, 1995.

COSTA, Lucas L. *et al.* **Alertas de fraude em licitações**: Uma abordagem baseada em redes sociais. In: Anais do XI Brazilian Workshop on Social Network Analysis and Mining. SBC, 2022. p. 37-48.

COSTALAT, Thallys Rubens Moreira; TAVARES, Géssica Fortes. **Machine learning techniques comparison for risk assessment of cardiovascular disease development by health indicators**. Brazilian Journal of Development, v. 8, n. 1, p. 6851-6862, 2022.

CUTLER, Adele; CUTLER, D. Richard; STEVENS, John R. ***Random Forest*s**. In: Ensemble machine learning. Springer, Boston, MA, 2012. p. 157-175.

DA COSTA, Carla Christina Ravaneda; DA VEIGA, Cássia Rita Pereira; DA VEIGA, Claudimar Pereira. **Experiência do consumidor e inteligência artificia**l: uma revisão da literatura. Desafio Online, v. 10, n. 3, 2022.

DAMACENO, Siuari Santos *et al.* **Inteligência artificial**: uma breve abordagem sobre seu conceito real e o conhecimento popular. Caderno de Graduação-Ciências Exatas e Tecnológicas-UNIT-SERGIPE, v. 5, n. 1, p. 11-11, 2018.

DANTAS, Adilmar Coelho *et al.* **AstroBot:** Um chatbot com inteligência artificial para auxiliar no processo de ensino e aprendizagem de física. In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. 2019.

DE CARVALHO JÚNIOR, Ciro Ferreira *et al.* **Chatbot**: uma visão geral sobre aplicações inteligentes. Revista Sítio Novo, v. 2, n. 2, p. 68-84, 2018.

DE CASTRO, F. C. C.; DE CASTRO, M. C. F. **Redes neurais artificiais**. Porto Alegre, RS: Pontifícia Universidade Católica do Rio Grande do Sul, 2001.

DOMINGOS, P. **A few useful things to know about machine learning**. Communications of the ACM, ACM, v. 55, n. 10, p. 78–87, 2012.

DOS SANTOS, Alba C. M. **A administração pública gerencial**. 2011. P. 15

ELKAN, Charles. The Foundations of Cost-Sensitive Learning. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle, 2001, p. 973-978.

FAWCETT, T. **An introduction to ROC analysis**. Pattern Recognition Letters, v. 27, n. 8, p. 861-874, 2006.

FÉLIX, Ana Catarina de Faria Blanc *et al.* **A influência da Inteligência Artificial No e-Commerce:** o Uso dos Chatbots. 2020. Tese de Doutorado. Universidade de Lisboa (Portugal).

FLACH, P. Automated Machine Learning: Methods, Systems, Challenges. Springer Press, 2014.

FLECK, Leandro *et al.* **Redes neurais artificiais**: Princípios básicos. Revista Eletrônica Científica Inovação e Tecnologia, v. 1, n. 13, p. 47-57, 2016.

FONSECA, Alberico Santos. **Fases de uma licitação**. 2014 Disponível em <http://www.fap-pb.edu.br/aluno/arquivos/material_didatico/direito/administrativo/fases_licitacao.pdf>. Acesso em 23 de novembro de 2022.

GARG, Pranav *et al.* **Learning invariants using decision trees and implication counterexamples.** ACM Sigplan Notices, v. 51, n. 1, p. 499-512, 2016.

GAREM, P. A. *et al.* Avaliação de métodos de aprendizado de máquina para avaliação de produtos óticos. Revista Brasileira de Orientação Profissional, v. 11, n. 2, p. 49-62, 2020.

GEMAN, S.; GEMAN, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. In: IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, v. 6, n. 6, p. 721-741, Nov. 1984.

GUIMARÃES, T. A. **A nova administração pública e a abordagem da competência**. 2011, Atlas. p. 44.

GURGEL, S.; FORMIGA, A. D. A. **Parallel implementation of feedforward neural networks on gpus**. In: Intelligent Systems (BRACIS), 2013 Brazilian Conference on. [S.l.: s.n.], 2013. p. 143–149.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. ***Random Forest*s**. In: The elements of statistical learning. Springer, New York, NY, 2009. p. 587-604.

HE, H.; GARCIA, E. A. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, v. 21, n. 9, p. 1263-1284, 2009.

HELLGE, C. *et al*. The curse of motivation in model selection: a Bayesian view. Statistics and Computing, v. 8, n. 3, p.187-198, 1998.

HU, Hanzhang *et al*. **Gradient boosting on stochastic data streams**. In: Artificial Intelligence and Statistics. PMLR, 2017. p. 595-603.

HUANG, H.; CHAWLA, Nitesh V.; BOWYER, Kevin W.; HALL, Lawrence O.; KEGELMEYER, W. **Validation of data mining results using sampling methods: A case study.** SIGKDD Explorations, v. 8, n. 2, p. 40-48, 2006.

ISHIKAWA, Lauro; DE ALENCAR, Alisson Carvalho. **Compliance inteligente**: o uso da inteligência artificial na integridade das contratações públicas. Revista de Informação Legislativa, v. 57, n. 225, p. 83-98, 2020.

JOHN, G.H.; KOHAVI, R.; PFLEGER, K. Irrelevant Features and the Subset Selection Problem. In: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 1997, pp. 121-129.

JORDAN, M. I. Learning in graphical models. Neural computation, v. 4, n. 1, p. 1-42, 1992.

KUHN, Max. **Building predictive models in R using the caret package**. Journal of Statistical Software, v. 28, n. 5, p. 1-26, 2008.

KOVÁCS, Zsolt László. **Redes neurais artificiais**. Editora Livraria da Fisica, 2002.

KOHAVI, R., JOHN, G.H., PFLEGER, K. Establishing a baseline in a lumpy environment. In: Proc. KDD. pp. 7–10. 1998.

KUHN, M. **Applied predictive modeling**. New York: Springer. 2014.

LANDGREBE, D.; DUIN, R. Feature extraction from Faces, p.91-111. 2ª ed. John Wiley & Sons, 2002.

LEUNG, K. Ming. **Naive bayesian classifier**. Polytechnic University Department of Computer Science/Finance and Risk Engineering, v. 2007, p. 123-156, 2007.

LEWIS, D. D.; CATLETT, C. Heterogeneous uncertainty sampling for supervised learning. In: Proceedings of the eleventh international conference on machine learning, p. 148-156, 1994.

LOPES, Marco Antonio. **Aplicação de Aprendizado de Máquina na Detecção de Fraudes Públicas.** Dissertação (Mestrado) - Universidade de São Paulo, 2019.

MADEIRA, Afonso Celso Magalhães; NEVES, Barbara Coelho; DANIEL DE JESUS, B. C. O **Uso da Inteligência Artificial Aplicada ao Marketing Digital**. Journal of Digital Media & Interaction, v. 3, n. 8, p. 95-111, 2020.

MATIAS, Ana Catarina de Faria Blanc Félix. **A influência da inteligência artificial no e-commerce:** o uso dos chatbots. 2020. Tese de Doutorado. Instituto Superior de Economia e Gestão.

MARTÍNEZ-GARCÍA, F. Aprendizaje máquina. Madrid, Espanha: Fundación para la Investigación y el Desarrollo Tecnológico de Asturias, 2012.

MAYER-SCHÖNBERGER, V.; CUKIER, K. **Big Data.** Edição traduzida. Rio de Janeiro: Elsevier, 2013.

MEIRELLES, Hely Lopes. **Licitação e Contrato Administrativo.** 5. ed. São Paulo: Malheiros Editores, 2004. p. 94.

MELLO, Celso Antônio Bandeira de. Curso de Direito Administrativo. 26. ed. São Paulo: Malheiros, 2009. p. 456

MIRANDA. Rodrigo Araújo de. **Gestão por Competências no Serviço Público**: o conhecimento como vantagem competitiva a Serviço da Administração Pública. Artigo Científico. Faculdade Projeção. 2010, p. 61

MITCHELL, T. Machine Learning. McGraw Hill, 1997.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. **Conceitos sobre aprendizado de máquina.** Sistemas inteligentes-Fundamentos e aplicações, v. 1, n. 1, p. 32, 2003.

MOREIRA, Diorginis Ormond; MIGNONI, Maria Eloisa. **Inteligência artificial**: o uso de chatbots no atendimento ao cliente. Revista Ibero-Americana de Ciências Ambientais, v. 12, n. 12, 2021.

MOTTA, Carlos Pinto Coelho. E**ficácia nas Licitações e Contratos:** comentários, doutrina e jurisprudência. 12. ed., rev. e atual. Belo Horizonte: Del Rey, 2011.

NASSIF, Ali Bou. **Short term power demand prediction using stochastic gradient boosting.** In: 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA). IEEE, 2016. p. 1-4.

NEVES, Bárbara Coelho. **Inteligência artificial e computação cognitiva em unidades de informação:** conceitos e experiências. Logeion: filosofia da informação, v. 7, n. 1, p. 186-205, 2020.

OLIVEIRA, Gilcelene Machado de. **Os princípios administrativos aplicáveis às licitações públicas:** as modalidades de licitação pública e o sistema de registro de preços. 2019. 60 f. Trabalho de Conclusão de Curso (Especialização em Gestão Pública Municipal) — Universidade de Brasília, Anápolis - GO, 2019.

PEDREGOSA, F. *et al.* **Scikit-learn**: Machine learning in Python. Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011.

PEREIRA, J. M. **Curso de Administração Pública**, 3. Ed. São Paulo: Atlas, 2010, p. 57.

PEREIRA, Luiz Carlos Bresser. **Da Administração Pública Burocrática à Gerencial**. Revista do Serviço Público, 47(1) janeiro-abril 2011, p. 12

PINTO, Vera Regina Ramos. **Um breve histórico sobre inovações em compras e licitações públicas no Brasil.** Brazilian Journal of Development, v. 6, n. 8, p. 63378-63397, 2020.

PORCHER, Bruno Casagrande. **Uso de chatbot como reforço para a aprendizagem de estudantes portadores de necessidades especiais**. MoExP-Mostra de Ensino, Extensão e Pesquisa do Campus Osório, v. 1, n. 1, p. 1-1, 2019.

POSTAL, Laura Casotti. **Robô no atendimento ao cliente:** quanto mais" humano" melhor? UFRGS, 2019.

RAUBER, Thomas Walter. **Redes neurais artificiais**. Universidade Federal do Espírito Santo, v. 29, 2005.

REVU, S.; SRIDHAR, A.; PADMAKUMAR, B. V. Understanding Model Accuracy - Precision, Recall, F1-score and Performance Measure Metrics. International Journal of Computer Science and Mobile Applications, v. 5, n. 2, p. 21-24, 2017.

RIJSSBERGEN, C. V. Information Retrieval. 2. ed. Butterworths, 1979.

RISH, Irina *et al.* **An empirical study of the naive Bayes classifier**. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. 2001. p. 41-46.

ROSILHO, André Janjácomo. **Qual é o modelo legal das licitações no Brasil?:** as reformas legislativas federais no sistema de contratações públicas. 2011. Tese de Doutorado.

R CORE TEAM (org.). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. [S. l.], 1997. Disponivel em: https://www.r-project.org/. Acesso em: 07 janeiro 2022.

RSTUDIO TEAM. **RStudio: integrated development for R.** Boston: RStudio, Inc, 2021. Disponível em: <https://www.rstudio.com/>. Acesso em: 07 janeiro 2022.

RUTZ, Samuel. **Licitação pública**. Revista de Direito e Economia da Concorrência, Volume 14, Edição 2, 2018.

SAMUEL, A. L. Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development, v. 3, n. 3, p. 210-229, jul. 1959.

SAMPAIO, Adilson Da Hora; FIGUEIREDO, Paulo Soares. **Análise de fragilidades no Sistema de Compras Públicas do Brasil:** verificação de indícios de Fraudes em Pregões Eletrônicos por meio da Lei de Benford. XLIII Encontro da ANPAD–EnANPAD, 2019.

SANTANNA, Felipe Alexandre, DANIEL, Anna Mucci. **A fase interna da licitação** - Distinções entre projeto básico e termo de referência. Disponível em <http://www.editoraforum.com.br/noticias/a-fase-interna-da-licitacao-distincoes-entre-projeto-basico-e-termo-de-referencia/>. Acesso em: 23 de novembro de 2022.

SCHNEIDER, Pedro Henrique. **Análise preditiva de Churn com ênfase em técnicas de Machine Learning:** uma revisão. 2016. Tese de Doutorado.

SERAFIM, D. C.; NETO, A. J. D. S. **Estruturando redes neurais artificiais paralelas e independentes para o controle de próteses robóticas**. Revista das Faculdades Integradas Claretianas, São Paulo, v. 6, dezembro 2013.

SETTLES, Burr. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009.

SINGH, S.; LEWIS, R. L.; KATRAK, H. Active learning for support vector machines. Journal of Machine Learning Research, v. 11, n. 10, p. 1471-1515, 2010.

SMOLA, A.; VISHWANATHAN, S. **Introduction to machine learning**. Cambridge University, UK, v. 32, p. 34, 2008.

Sokolova, M., Japkowicz, N., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. 45 (4), 427–437.

SOUZA, Solange Seléto; MAURO, Maria Yvone Chaves. **O uso da robótica, games, realidade virtual e realidade aumentada no tratamento de autismo, demência, esquizofrenia e fobia.** CADERNO DE PESQUISA APLICADA, v. 1, n. 3, p. 1-11, 2021.

SOUTO, D.; REALI, A.; NASCIMENTO, A. C. do; SILVA, F. R. da. A comparison of machine learning classification methods applied to malaria diagnostics. Expert Systems with Applications, v. 56, p. 37-47, 2016.

SUTTON, R. S. et al. Aprendizado por Reforço. In: Inteligência Artificial. Rio de Janeiro: Elsevier, 2018.

SUN, Y.; SHEKHAR, S.; CHAWLA, Nitesh V. **Mining Imbalanced Data Sets: An Overview.** In: IEEE International Conference on Data Mining, 2002, Maebashi. Proceedings of the 2002 IEEE International Conference on Data Mining, 2002.

TEIXEIRA, A. Governar Inovando. **Os desafios contemporâneos da Gestão Pública**. Revista Poder Local No. 1. Fortaleza: Omni editora, 2011, p. 51

TEIXEIRA, Alberto. Governar Inovando. **Os desafios contemporâneos da Gestão Pública.** Revista Poder Local No. 1. Fortaleza: Omni editora, 2011, p. 51

TIBESHIRAI, R. et al. The Elements of Statistical Learning. New York: Springer, 2003.

TRIBUNAL DE CONTAS DA UNIÃO. Disponível em <https://portal.tcu.gov.br/início>. Acesso em 23 de novembro de 2022.

WEBB, Geoffrey I.; KEOGH, Eamonn; MIIKKULAINEN, Risto. **Naïve Bayes**. Encyclopedia of machine learning, v. 15, p. 713-714, 2010.

XU, Shuo. **Bayesian Naïve Bayes classifiers to text classification**. Journal of Information Science, v. 44, n. 1, p. 48-59, 2018.

YAMADA, I.; ROELOFS, L. M.; HALL, G. P. Detecting fraud in public procurement using data mining. In: 2016 IEEE International Conference on Big Data (BigData 2016), p. 2079–2083. 2016.

ZAREPOUR, M.; HOSSEINI, S. A.; NAFIAN, B.; ZAREFAHAND, M.; TESHNEHLAB, M. Abyaneh. **Natural Landmark Recognition through Machine Learning Algorithms**. International Journal of Computer Applications, v. 137, n. 7, p. 1-6, 2016.

ZHANG, K.; ZHOU, Z.; HEYWOOD, M. ROC Analysis for Classifiers with Unbalanced Error Costs. Journal of Machine Learning Research, v. 5, p. 1057-1069, 2004.

LAPEDES, A.S., FARBER, R.M. **Nonlinear signal processing using neural networks: prediction and system modeling.** 1987.

SHALEV-SHWARTZ, S., SINGER, Y., SREBRO, N., Ben-David, S. **Understanding machine learning: From theory to algorithms**. Cambridge University Press, Cambridge. 2014

APÊNDICE A – Matrizes de Confusão

- **Regressão Logística**

Quadro – Classificações do modelo

		Valor predito	
		Sim	Não
Valor Real	Sim	96900	70381
	Não	19638	45711

Fonte: Elaboração própria, 2023

- ***Random Forest***

Quadro – Classificações do modelo

		Valor predito	
		Sim	Não
Valor Real	Sim	114688	1850
	Não	9265	106827

Fonte: Elaboração própria, 2023

- **Redes neurais**

Quadro – Classificações do modelo

		Valor predito	
		Sim	Não
Valor Real	Sim	99427	17111
	Não	75096	40996

Fonte: Elaboração própria, 2023

- ***Naïve Bayes***

Quadro – Classificações do modelo

		Valor predito	
		Sim	Não
Valor Real	Sim	101712	14826
	Não	76779	39313

Fonte: Elaboração própria, 2023

- ***Stochastic Gradient Boosting***

Quadro – Classificações do modelo

		Valor predito	
		Sim	Não
Valor Real	Sim	99427	17111
	Não	75096	40996

Fonte: Elaboração própria, 2023

- **Árvore de Decisão**

Quadro – Classificações do modelo

		Valor predito	
		Sim	Não
Valor Real	Sim	99427	17111
	Não	75096	40996

Fonte: Elaboração própria, 2023

APÊNDICE B - SCRIPT PARA COLETA DE DADOS

```
library(httr)

# http://compras.dados.gov.br/docs/home.html
# https://www.gov.br/compras/pt-br/images/ultimas_noticias/lista-uasgsa_sisg.pdf

#link<-'http://compras.dados.gov.br/contratos/v1/contratos.csv?offset=500'
#link<-'http://compras.dados.gov.br/contratos/v1/contratos.csv'

#dados<-httr::get(link,encode = "csv")
#dados2<-content(dados)
#rbind(dados2,dados2)

banco<-c()
for(i in seq(528500,900000,500)){
  link<-paste0('http://compras.dados.gov.br/contratos/v1/contratos.csv?offset=',i)
  dados<-httr::get(link,encode = "csv")
  dados2<-content(dados)
  banco<-rbind(banco,dados2)
  cat("\r", i, "of", 600000)
)

save(banco1,file="c:/users/hp/desktop/tccunirio/gabriel/banco_gabriel_parte8.rdata")
```

APÊNDICE C – SCRIPT DE TRATAMENTO DE DADOS

```
# banco de contratos
banco_completo <- readRDS("C:/Users/gabri/OneDrive/Área de Trabalho/TCC_Machine Learning/dados_recentes/banco_completo.Rds")

names(banco_completo)

# selecionar as variaveis importantes
banco_completo <-banco_completo[,c(1:3,7:10,12:18)]

names(banco_completo)

# banco CEIS
library(readr)

library(janitor)

banco_CEIS <- read_delim("C:/Users/gabri/OneDrive/Área de Trabalho/TCC_Machine Learning/dados_recentes/20220107_CEIS.csv",
                         ";" ,escape_double = FALSE, trim_ws = TRUE) %>% clean_names()

# selecionar as variaveis importantes
banco_CEIS <- banco_CEIS[,c(1:5,7:14,17:19)]

# separar o CPF e CNPJ da base CEIS
library(dplyr)

banco_CEIS_PJ <- banco_CEIS %>% filter(tipo_de_pessoa=="J")

banco_CEIS_PF <- banco_CEIS %>% filter(tipo_de_pessoa=="F")

remove(banco_CEIS)
```

```
# separar o CPF e o CNPJ da base completa
banco_completo_PJ <-banco_completo %>% drop_na(`CNPJ Contratada`)
banco_completo_PF <-banco_completo %>% drop_na(`CPF Contratada`)

remove(banco_completo)

#CNPJ CEIS = 92136704000104
#CNPJ CEIS = 92.136.704/0001-04
#CNPJ CONTRATOS = Fornecedor 03.701.471/0001-15: ORGAL VIGILANCIA E SEGURANCA LTDA

# jogando fora a palavra "Fornecedor"
banco_completo_PJ$cnpj <- banco_completo_PJ$`CNPJ Contratada`
banco_completo_PJ$cnpj <- gsub("Fornecedor","",banco_completo_PJ$cnpj)

# dividindo o cnpj e  o nome da empresa
library(tidyr)

banco_completo_PJ<-banco_completo_PJ %>% separate(cnpj, c("cnpj", "nome"),sep=":")

# tirando o "." o "-" e a "/"
banco_completo_PJ$cnpj <- gsub("\\.","",banco_completo_PJ$cnpj)
banco_completo_PJ$cnpj <- gsub("\\-","",banco_completo_PJ$cnpj)
banco_completo_PJ$cnpj <- gsub("\\/","",banco_completo_PJ$cnpj)

banco_CEIS_PJ <-rename(banco_CEIS_PJ, cnpj = cpf_ou_cnpj_do_sancionado)

#verificando a classe
class(banco_completo_PJ$cnpj)
class(banco_CEIS_PJ$cnpj)


banco_completo_PJ$cnpj <- gsub("\\W", "", banco_completo_PJ$cnpj)
```

```
# juntando as duas bases de dados

BANCO <- banco_completo_PJ %>% left_join(banco_CEIS_PJ)
table(BANCO$tipo_de_pessoa)

#Empresas Inidôneas e Suspensas - EIS
BANCO$EIS <- ifelse(BANCO$tipo_de_pessoa=="J",1,0)
BANCO$EIS <- ifelse(is.na(BANCO$EIS),0,BANCO$EIS)

table(BANCO$EIS)

remove(banco_CEIS_PF,banco_CEIS_PJ,banco_completo_PF,banco_completo_PJ)

# tratamento da variável "valor"
BANCO <-rename(BANCO, valor = `Valor inicial`)
BANCO$valor <- gsub("R\\$", "", BANCO$valor)
BANCO$valor <- gsub("\\.", "", BANCO$valor)
BANCO$valor <- gsub("\\,", ".", BANCO$valor)
BANCO$valor <-as.double(BANCO$valor)

summary(BANCO$valor)


BANCO <-rename(BANCO,aditivos =`Número de Aditivos`)
BANCO$aditivos<-as.numeric(BANCO$aditivos)
BANCO$aditivos <- ifelse(is.na(BANCO$aditivos),0,BANCO$aditivos)
summary(BANCO$aditivos)
BANCO <-rename(BANCO,modalidade =`Modalidade da Licitação`)

BANCO<-BANCO %>% separate(UASG, c("num_uasg", "uasg"),sep=":")

BANCO$uasg2 <-ifelse(BANCO$num_uasg>15000, 'EDUCACAO',
```

```
        ifelse(BANCO$num_uasg>13000, 'AGRICULTURA',
        ifelse(BANCO$num_uasg>12000, 'MILITAR',
        ifelse(BANCO$num_uasg>110096,'ADMINISTRACAO',
        ifelse(BANCO$num_uasg>110062,'PROCURADORIA',
        'OUTROS')))))

table(banco$uasg2)

table(banco$modalidade)
```

APÊNDICE D – SCRIPT DE APLICAÇÃO DE ALGORITMOS DE ML

```
# loading the database
BANCO <- readRDS("C:/Users/suely/Desktop/modelos e banco/banco_tratado.Rds")

# display the column names
names(BANCO)

# oversampling the minority class
library(ROSE)

banco <- ROSE::ovun.sample(formula = EIS ~ valor+aditivos+modalidade+uasg2, data = BANCO, p=0.5, seed = 1, method = "over")$data

# display the class distribution of the original and oversampled datasets
table(BANCO$EIS)

table(banco$EIS)

rm(BANCO)

# convert 'EIS', 'modalidade','uasg2' to a factor variable

banco$EIS<-as.factor(banco$EIS)

banco$modalidade<-as.factor(banco$modalidade)

banco$uasg2<-as.factor(banco$uasg2)

#-------------------------------------------------------------------------------
# Train and Test
#-------------------------------------------------------------------------------
# set seed for reproducibility
```

```
set.seed(12345)
dim(banco)

# split dataset into train and test
smp_size <- floor(0.8 * nrow(banco))

train_ind <- sample(seq_len(nrow(banco)), size = smp_size)

train <- banco[train_ind, ]
test  <- banco[-train_ind, ]
train<-na.omit(train)
test<-na.omit(test)

#-------------------------------------------------------------------------------------
# Modelo 1 - logit
#-------------------------------------------------------------------------------------

# fit logistic regression model
library(broom)
modelo1  <- glm(EIS ~ valor+aditivos+modalidade+uasg2 , family =

                binomial(link = "logit"), data = train)

# get the summary of the model
summary(modelo1)

# make predictions on the test dataset
yhat <- predict(modelo1, newdata=test, type = "response")

# create a dataframe with the predictions
predictions <- data.frame(yhat, EIS = test$EIS)

# give a more meaningful name to the column with predictions
colnames(predictions)[1] <- "Predicted_Prob"
```

```
# calculate classification threshold and make binary predictions
predictions$Predicted_Class <- ifelse(predictions$Predicted_Prob
                                      > 0.5, 1, 0)

predictions$Predicted_Class<-as.factor(predictions$Predicted_Class)

predictions$EIS <- as.factor(predictions$EIS)

# evaluate model performance
library(caret)

metrics <- confusionMatrix(predictions$EIS, predictions$Predicted_Class,
                           mode = "everything", positive = "1")
metrics

#------------------------------------------------------------------------
#Calculation of variable importance for regression and classification models
#A generic method for calculating variable importance for objects produced
#by train and method specific methods
#------------------------------------------------------------------------

# obter os coeficientes do modelo
coefs <- coef(modelo1)

# imprimir a importância das variáveis
print(coefs)

library(caret)
giniImp <- varImp(modelo1, scale = TRUE)
giniImp

#------------------------------------------------------------------------
# colinearidade
```

```
#------------------------------------------------------------------------
car::vif(modelo1)

#------------------------------------------------------------------------

library(ggplot2)
library(ROCR)

predictions$Predicted_Class <- as.numeric(predictions$Predicted_Class) -1

pred <- prediction(predictions$Predicted_Class, predictions$EIS)
perf <- performance(pred, measure = "tpr", x.measure = "fpr")



ggplot() +
  geom_line(aes(x = perf@x.values[[1]], y = perf@y.values[[1]])) +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed") +
  ggtitle("ROC curve") + xlab("False Positive Rate") + ylab("True Positive Rate")+
  geom_text(x=0.6, y=0.2, label = paste("AUC = ", round(performance(pred,
"auc")@y.values[[1]], 2)))

#Alternative

library(pROC)
roc_obj <- roc(response = predictions$EIS, predictor = predictions$Predicted_Class)
plot(roc_obj, print.auc = TRUE)

rm(giniImp, metrics, modelo1, perf, pred, predictions, yhat, banco)
#-----------------------------------------------------------------------------------
# Modelo 2 - Random Forest
#-----------------------------------------------------------------------------------
library(caret)
library(ranger)
```

```
# Definir a validação cruzada com 5 dobras
trctrl <- trainControl (method = "cv",number=5)

# Definir o treinamento com 10 dobras de validação cruzada, repetido 3 vezes
trctrl <- trainControl(method='repeatedcv',
                       number=10,
                       repeats=3)


# Treinar o modelo de floresta aleat?ria
set.seed(123)

RF <- train(EIS ~ ., data = train,
            method = "ranger",
            trControl = trctrl,
            tuneGrid = expand.grid(mtry = c(1:ncol(train)),
                                   splitrule = c("gini", "extratrees"),
                                   min.node.size = c(1,5,10)),
            importance = "impurity")


# Salvar o modelo treinado
save(RF,file = 'C:/Users/gabri/OneDrive/?rea de Trabalho/TCC_Machine Learning/dados_recentes/RF2.Data')

# Carregar o modelo treinado
load(file = 'C:/Users/suely/Desktop/modelos e banco/RF2.Data')

# Make predictions on the test set
yhat <- predict(RF, newdata = test)

# Create dataframe with predicted outcome and true outcome

RF_predicted <- data.frame(yhat = yhat, EIS = test$EIS)
```

```
RF_predicted$yhat <- as.factor(RF_predicted$yhat)

RF_predicted$EIS <- as.factor(RF_predicted$EIS)

# Compare predicted outcome and true outcome
table(RF_predicted$yhat, RF_predicted$EIS)

# Calculate confusion matrix
confusionMatrix(RF_predicted$yhat, RF_predicted$EIS, mode = "everything", positive = "1")

#Import?ncia das vari?veis

varImp (RF$finalModel)
varImpPlot (RF, sort = TRUE, n.var = 10)
print(summary(RF))

# plotando o gr?fico ROC
plot(RF, type = "roc")

# removendo vari?veis
rm('yhat','RF','RF_predicted')

#-------------------------------------------------------------------------------------
#  Modelo 3 - REDES NEURAIS
# we will now train neural net model
#-------------------------------------------------------------------------------------

library(caret)
rede_neural <- train(EIS~.,
                     data = train,
                     method = "nnet",
                     trControl=trctrl,
```

```
                # this is maximum number of weights
                # needed for the nnet method

                MaxNWts=2000)
```

save(rede_neural,file = 'C:/Users/gabri/OneDrive/?rea de Trabalho/TCC_Machine Learning/dados_recentes/modelo_rede_neural2.Data')

load(file = 'C:/Users/suely/Desktop/modelos e banco/modelo_rede_neural2.Data')

```
#------------------------------------------------------------------------------------
varImp(rede_neural, scale=FALSE)
plot(varImp(rede_neural, scale=FALSE))
#------------------------------------------------------------------------------------

yhat=predict(rede_neural,newdata=test)
predicted <- data.frame(yhat)
predicted$EIS<-test$EIS

#------------------------------------------------------------------------------------
predicted$yhat <- as.factor(predicted$yhat)
predicted$EIS <- as.factor(predicted$EIS)

confusionMatrix(predicted$yhat,predicted$EIS,mode = "everything", positive = "1")

rm('predicted','rede_neural','yhat')

#------------------------------------------------------------------------------------
# Modelo 4 - Naive Bayes
#------------------------------------------------------------------------------------
library(e1071)
library(caret)
```

```
modelo4<-naiveBayes(EIS ~ valor+aditivos+modalidade+uasg2 , data = train)
summary(modelo4)
modelo4

save(modelo4,file = 'C:/Users/gabri/OneDrive/?rea de Trabalho/TCC_Machine Learning/dados_recentes/modelo_naiveBayes2.Data')

load(file = 'C:/Users/suely/Desktop/modelos e banco/modelo_naiveBayes2.Data')

yhat4=predict(modelo4,newdata=test)
predicted4 <- data.frame(yhat4)
predicted4$EIS <-test$EIS


predicted4$yhat4 <- as.factor(predicted4$yhat4)
predicted4$EIS <- as.factor(predicted4$EIS)

confusionMatrix(predicted4$yhat4,predicted4$EIS, mode = "everything", positive = "1")

varImp(modelo4)

remove('yhat4','modelo4','predicted4')

#-------------------------------------------------------------------------------
## Modelo 5 - Stochastic Gradient Boosting
#-------------------------------------------------------------------------------

Stochastic_Gradient_Boosting <- train(EIS ~ ., data = train,
                                      method = "gbm",
                                      trControl = trctrl,
                                      ## This last option is actually one
                                      ## for gbm() that passes through
```

```
                                verbose = TRUE)

        save(Stochastic_Gradient_Boosting,file = 'C:/Users/gabri/OneDrive/?rea de
Trabalho/TCC_Machine
Learning/dados_recentes/modelo_Stochastic_Gradient_Boosting2.Data')

        load(file = 'C:/Users/suely/Desktop/modelos e
banco/modelo_Stochastic_Gradient_Boosting2.Data')

        yhat=predict(Stochastic_Gradient_Boosting,newdata=test)
        predicted <- data.frame(yhat)
        predicted$EIS<-test$EIS

        predicted$yhat <- as.factor(predicted$yhat)
        predicted$EIS <- as.factor(predicted$EIS)

        confusionMatrix(predicted$yhat,predicted$EIS, mode = "everything", positive = "1")

        varImp(Stochastic_Gradient_Boosting)

        plot(varImp(Stochastic_Gradient_Boosting), scale(TRUE),top=10)

        rm(predicted,Stochastic_Gradient_Boosting,yhat)

        #-------------------------------------------------------------------------------
        # Modelo 6 - Decision tree
        #-------------------------------------------------------------------------------
        # load libraries
        library(rpart)
        library(rattle)

        rpart <- rpart(EIS ~ ., data=train, method="class")
        rpart
```

```
save(rpart,file = 'C:/Users/gabri/OneDrive/?rea de Trabalho/TCC_Machine Learning/dados_recentes/modelo_decision_tree2.Data')

load(file = 'C:/Users/suely/Desktop/modelos e banco/modelo_decision_tree2.Data')

# plot decision tree

fancyRpartPlot(rpart, main="ss")
plot(rpart)
text(rpart,pretty=0)

yhat=predict(rpart,newdata=test)
predicted <- data.frame(yhat)
predicted$yhat=ifelse(predicted$X1>=0.5, 1, 0)
predicted$EIS<- test$EIS

predicted$yhat <- factor(predicted$yhat, levels=c(0, 1), ordered=TRUE)
predicted$EIS <- as.factor(predicted$EIS)

confusionMatrix(predicted$yhat,predicted$EIS, mode = "everything", positive = "1")

# AUC e ROC
library(pROC)
predicted$yhat <- factor(predicted$yhat, levels=c(0, 1), ordered=TRUE)
predicted$EIS <- factor(predicted$EIS, levels=c(0, 1), ordered=TRUE)
auc <- roc(predicted$EIS, predicted$yhat)
auc
plot(auc)

remove('predicted','rpart','yhat','auc')

#-------------------------------------------------------------------------------
# Modelo 3 - SVM - não funcionou
#-------------------------------------------------------------------------------
```

```
# Fit the model
svm3 <- train(EIS~.,data = train, method = "svmRadial", trControl = trctrl, preProcess = c("center","scale"), tuneLength = 10,
              maxit = 10,verbose = TRUE)

#-------------------------------------------------------------------------------
# Modelo 5 - KNN - não funcionou
#-------------------------------------------------------------------------------

library(class)
previsoes <- knn(train = train[,-4], test= test[,-4],cl= train[,4],k=1)
head(previsoes)
```