Instituto Serzedello Corrêa – ISC
Pós-Graduação em Análise de dados para o Controle

Thiago Menegardo Nunes

Classificação de editais de pregões eletrônicos por meio de algoritmos tradicionais de aprendizagem de máquina

Brasília
2020

Thiago Menegardo Nunes

Classificação de editais de pregões eletrônicos por meio de algoritmos tradicionais de aprendizagem de máquina

Trabalho de conclusão de curso submetido ao Instituto Serzedello Corrêa do Tribunal de Contas da União como requisito para a obtenção do grau de especialista.
Orientador: Prof. Dr. Eduardo Chaves Ferreira

Brasília
2020

Thiago Menegardo Nunes

Classificação de editais de pregões eletrônicos por meio de algoritmos tradicionais de aprendizagem de máquina

Trabalho de conclusão do curso de pós-graduação realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista.

Brasília, 30 de março de 2020.

Banca Examinadora:

_____

Prof. Dr. Eduardo Chaves Ferreira

Orientador

_____

Sylvio Xavier Júnior

Examinador

Resumo

A Secretaria de Fiscalização de Tecnologia da Informação do Tribunal de Contas da União é a unidade técnica responsável pela fiscalização das licitações sobre tecnologia da Informação de todos os órgãos e entidades da Administração Pública. Dessa forma, a avaliação tempestiva e automatizada de editais de licitação com a identificação de indícios de irregularidades, fraudes, desvios e desperdícios de recursos públicos possibilita ações de controle mais eficientes e efetivas. Este trabalho tem como o objetivo o desenvolvimento de um algoritmo que disponibilize, automaticamente, informações sobre editais de licitações, resultados de pregões e extratos de dispensa e inexigibilidade de licitação cujo tema seja tecnologia da informação. O modelo desenvolvido neste trabalho apresentou resultados satisfatórios.

Palavras-chave: Aprendizagem de máquina; Classificação textual; Random Forest Classifier.

Abstract

Information Technology Audit Department is a technical department responsible for the review of the information technology government acquisition process of all organizations of the federal government. Thus, the automated and timely evaluation of a bidding process and the identification of signs of law-breaking, illicit act, bid rigging, embezzlement of public funds and misuse of public money enables the selection of efficient and effective audit topics. The purpose of this study is to build a machine learning algorithm capable of provide information about government acquisition process, such as, sole-source acquisition and no-bid contract, whose subject is information technology. The text classifier built in this study produced satisfactory results.

Keywords: Machine learning; Text Classification; Random Forest Classifier

Sumário

1. Introdução

A avaliação tempestiva e automatizada de editais de licitação e atas de pregão, com a identificação de indícios de irregularidades, fraudes, desvios e desperdícios de recursos públicos, possibilita ações de controle mais eficientes e efetivas por parte do Tribunal de Contas da União.

Considerando que a Secretaria de Fiscalização de Tecnologia da Informação (Sefti) é a unidade técnica do Tribunal responsável pela fiscalização das licitações sobre tecnologia da Informação de todos os órgãos e entidades da Administração Pública, a seleção de licitações por área temática, como, por exemplo, tecnologia da informação, torna-se imprescindível.

Este trabalho visa, por meio de técnicas de classificação textual, disponibilizar à Sefti, informações sobre editais de licitações, resultados de pregões e extratos de dispensa e inexigibilidade de licitação cujo tema seja tecnologia da informação e que estão disponíveis nos sistemas Siasg e Comprasnet. Espera-se que o resultado permita a categorização dos editais de licitação possibilitando à Sefti uma atuação mais efetiva na fiscalização dos gastos públicos em TI.

Classificação textual é o processo de atribuir um conjunto de categorias preestabelecidas a um texto ou documento de acordo com o seu conteúdo. Ou seja, um classificador textual recebe um texto como entrada, analisa seu conteúdo, e atribui automaticamente categorias, como, por exemplo, tecnologia da informação, obras, etc.

Neste trabalho foi utilizada a metodologia CRISP-DM, um modelo de processo de mineração de dados e que consiste de seis etapas: entendimento do negócio; entendimento dos dados; preparação dos dados; modelagem; avaliação; e implantação do modelo.

2. Problema e justificativa

O corpo técnico do Tribunal de Contas da União dispõe do Sistema de Análise de Licitações e Editais (ALICE), ferramenta que possibilita a avaliação tempestiva e automatizada de editais de licitação e atas de pregão, com a identificação de indícios de irregularidades, fraudes, desvios e desperdícios de recursos públicos, possibilitando ações de controle mais eficientes e efetivas.

Segundo o documento “Sistemática de análise das informações fornecidas por meio dos e-mails diários do sistema Alice para as unidades técnicas do Tribunal de Contas da União”, publicado no Boletim do Tribunal de Contas da União em 22/10/2018, esse sistema disponibiliza, às unidades técnicas do Tribunal, por meio de três e-mails diários, informações sobre editais de licitações, resultados de pregões e extratos de dispensa e inexigibilidade de licitação. Cada unidade técnica recebe apenas e-mails com informações das aquisições federais publicadas relativas à sua clientela, ou seja, às Uasgs (Unidades Administrativas de Serviços Gerais integrantes do Siasg – Sistema Auxiliar de Serviços Gerais) de sua responsabilidade.

A Secretaria de Fiscalização de Tecnologia da Informação (Sefti), uma das unidades técnicas do Tribunal, possui a finalidade de fiscalizar a gestão e o uso de recursos de tecnologia da informação pela Administração Pública Federal, ou seja, é responsável por fiscalizar as licitações sobre tecnologia da Informação de todas as Uasgs. Dessa forma, a Sefti recebe diariamente e-mails do Alice sobre licitações cujo tema seja tecnologia da informação, uma vez que esse sistema propicia a seleção de pregões por área temática, como obras e TI.

No entanto, essa seleção por área temática gera um alto número de falsos positivos, o que atrapalha o acompanhamento sistemático e tempestivo das aquisições na área de TI realizadas com recursos públicos federais.

Assim, a justificativa do presente trabalho é a necessidade de se desenvolver um classificador de editais de licitação capaz de selecionar apenas as licitações cujo tema seja TI sem, no entanto, gerar um alto número de falsos positivos e que, também, possua um número baixo de falsos negativos.

3. Objetivos

A partir da definição do problema a ser tratado e do escopo de pesquisa, definem-se os objetivos, classificando-os em:

3.1. Objetivo geral

Utilizar técnicas tradicionais para desenvolver um modelo com alto recall e alta precisão para classificar editais de licitação do tipo pregão eletrônico a partir de treinamento realizado com licitações obtidas do BD_SIASG e selecionadas a partir de empenhos com natureza de despesa relacionada com objetos de TI. Depois, utilizar o modelo treinado para classificar os novos editais presentes no sistema Comprasnet e enviar e-mails à Sefti contendo as licitações cujo tema seja TI.

3.2 Objetivos específicos

- Classificar os editais dos pregões eletrônicos cadastrados no BD_COMPRASNET;
- Assegurar alto grau de recall e de precisão;
- Facilitar a conversão do protótipo em solução, a ser definida por meio de projeto futuro.

4. Metodologia

A metodologia utilizada neste trabalho foi a CRISP-DM (Cross Industry Standard Process for Data Mining), uma metodologia de mineração de dados em formato cíclico, composta por seis fases, que direcionam a descoberta do conhecimento para tomada de decisão sobre dados em grande volume. O fluxo entre as fases é não unidirecional, possibilitando ir e voltar entre as suas fases e tarefas (Chapman et al, 2000).

As fases da metodologia CRISP-DM são: Entendimento do Negócio; Entendimento dos Dados; Preparação dos Dados; Modelagem; Avaliação; e Implementação do Modelo (Chapman et al, 2000).

A fase de entendimento do negócio se concentra no entendimento dos objetivos e requisitos do projeto sob uma perspectiva de negócios, convertendo esse conhecimento em uma definição de problema de mineração de dados e em um plano preliminar projetado para atingir os objetivos (Chapman et al, 2000).

A fase de entendimento dos dados começa com a coleta inicial de dados e prossegue com as atividades que permitem que você se familiarize com os dados, identifique problemas de qualidade dos dados, descubra as primeiras idéias sobre os dados ou detecte subconjuntos interessantes para formar hipóteses sobre informações ocultas (Chapman et al, 2000).

A fase de preparação dos dados abrange todas as atividades necessárias para construir o conjunto de dados final a partir dos dados brutos iniciais. É provável que as tarefas de preparação de dados sejam executadas várias vezes e não em qualquer ordem prescrita. As tarefas incluem seleção de tabela, registro e atributo, além de transformação e limpeza de dados para ferramentas de modelagem (Chapman et al, 2000).

Na fase da modelagem, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para otimizar valores. Normalmente, existem várias técnicas para o mesmo tipo de problema de mineração de dados (Chapman et al, 2000).

Na fase da avaliação, você construiu um modelo que parece ter alta qualidade a partir de uma perspectiva da análise de dados. Antes de prosseguir

para a implantação final do modelo, é importante o avaliar completamente e revisar as etapas executadas para sua criação, para garantir que o modelo atinja adequadamente os objetivos de negócios. Um objetivo chave é determinar se há algum problema comercial importante que não foi suficientemente considerado. No final desta fase, uma decisão sobre o uso dos resultados da mineração de dados deve ser alcançada (Chapman et al, 2000).

Na fase de implementação do modelo o conhecimento adquirido precisará ser organizado e apresentado de forma que o cliente o possa usar (Chapman et al, 2000).

Dessa forma, o desenvolvimento do presente trabalho será descrito nas seções a seguir, uma para cada fase da metodologia CRISP-DM.

5. Entendimento do negócio

A fase de entendimento do negócio é composta pelas seguintes tarefas: i) determinar os objetivos do negócio, que busca entender, da perspectiva do negócio, o problema a ser solucionado; ii) avaliar a situação, que envolve uma investigação mais detalhada sobre todos os recursos, restrições, premissas e outros fatores que devem ser considerados na determinação da meta de análise de dados e do plano do projeto; iii) identificar os objetivos da mineração de dados, em que os critérios para se alcançar os objetivos de negócio serão descritos em termos técnicos; e iv) produzir o plano de projeto, contendo os passos necessários para atingir as metas de mineração de dados e, assim, atingir as metas de negócio (Chapman et al, 2000).

5.1 Objetivo do Negócio

Embora se possa dizer que o sistema adotado pelo Brasil seja o chamado controle “a posteriori”, tecnicamente o TCU o exerce, em certas circunstâncias, “concomitantemente”, quando exercita auditorias, inspeções e acompanhamentos, e até “previamente”, ao examinar, por exemplo, os editais de licitações públicas (Nagel, 2015).

O modelo de controle ‘a posteriori’ sofreu grande expansão, mas não é hoje o adotado de forma exclusiva por muitos países, uma vez que sua utilidade é pequena e possui elevado custo para os órgãos de controle. Por último, podemos destacar que o modelo de controle concomitante é atualmente a fórmula encontrada pelos países para modernizar suas instituições (Tribunais ou Controladorias), dando um caráter ágil e eficiente ao exercício da fiscalização. Este sistema, que convive em muitos órgãos com formas de controle posterior e prévio, é hoje o que melhor se adapta às necessidades do Estado Moderno, permitindo aos órgãos de controle exercitar as atualizadas técnicas de auditoria e atender aos reclamos do Parlamento e da opinião pública” (CITADINI, 1995).

A Sefti é a unidade do TCU responsável por fiscalizar a gestão e o uso de recursos de TI na Administração Pública Federal. Essa fiscalização é realizada, por exemplo, por meio de acompanhamentos, levantamentos, inspeções, auditorias ou monitoramentos. Ademais, possui como estratégia exercer o controle sobre os

pregões eletrônicos à medida que os atos ou atividades são executados, objetivando a adoção de medidas saneadoras, como o intuito de evitar que seja realizado apenas o controle corretivo, exercido após a conclusão do objeto, visando a responsabilização dos gestores e a reparação do dano.

Uma das formas utilizadas pela Sefti para exercer o controle prévio e concomitante da gestão e do uso de recursos de TI é por meio de fiscalização do tipo Acompanhamento, que é o instrumento de fiscalização utilizado pelo Tribunal para examinar a legalidade e a legitimidade dos atos de gestão dos responsáveis sujeitos à sua jurisdição, quanto ao aspecto contábil, financeiro, orçamentário e patrimonial, e para avaliar o desempenho dos órgãos e entidades jurisdicionadas quanto aos aspectos de economicidade, eficiência e eficácia dos atos praticados, por um período predeterminado.

Assim, as atividades dos órgãos e entidades são acompanhadas de forma seletiva e concomitante, mediante informações obtidas dos editais de licitação, consulta a sistemas informatizados adotados pela APF, acesso a informações publicadas em sítio eletrônico do órgão ou unidade, entre outros.

Dessa forma, e em virtude da identificação, nos últimos anos, de uma série de licitações com falhas no planejamento, muitas vezes direcionadas e com referenciais ruins de preços, que se transformam em contratos com elevado sobrepreço ou superfaturamento, a Sefti, alinhada ao objetivo estratégico do Plano Estratégico do Tribunal de Contas da União 2019-2025, de contribuir para melhorar a capacidade de contratação das organizações públicas, decidiu realizar um acompanhamento de aquisições na área de TI promovidas por órgãos e entidades da APF, com o objetivo de atuar de forma preventiva, antes da ocorrência do pregão ou da contratação, uma vez que é a forma mais efetiva de evitar prejuízos que dificilmente seriam recuperados com uma ação a posteriori.

As informações obtidas para esse acompanhamento tiveram como fonte ferramentas como os painéis Alice, Adele e Mônica, instrumentos que permitem saber, quase concomitantemente, quais licitações estão ocorrendo. No entanto, as informações sobre licitações cujo tema seja tecnologia da informação são geradas com um alto número de falsos positivos, o que atrapalha o acompanhamento

sistemático e tempestivo das aquisições na área de TI realizadas com recursos públicos federais.

O Alice (Análise de Licitações e Editais), é uma ferramenta que possibilita a avaliação tempestiva e automatizada de editais de licitação e atas de pregão, com a identificação de indícios de irregularidades, fraudes, desvios e desperdícios de recursos públicos, possibilitando ações de controle mais eficientes e efetivas.

O painel MONICA (Monitoramento Integrado para o Controle de Aquisições) apresenta informações relativas às aquisições efetuadas pela esfera federal, incluindo os poderes Executivo, Legislativo e Judiciário, além do Ministério Público. No entanto, não constam do painel as aquisições efetuadas por meio de sistemas diversos do Siasg (normalmente efetuadas por estatais) e também aquelas processadas por meio do Regime Diferenciado de Contratações (RDC). Os dados obtidos por meio do Comprasnet (pregões) são atualizados semanalmente, ao passo que as demais informações são carregadas mensalmente.

O painel Adele (Análise da disputa em licitações eletrônicas) é uma ferramenta de tecnologia da informação que foi idealizada com o intuito de melhorar as análises relativas ao nível de competitividade em certames licitatórios. Por meio do painel, é possível obter, graficamente, informações objetivas sobre a competição (ou não) em determinado pregão. Além do gráfico de competição, a Adele segmenta as informações por fase do pregão (propostas, lances e aleatória) e apresenta dados resumidos e analíticos. O Adele apenas exibe informações de pregões processados por meio do Comprasnet.

Dessa forma, a Sefti necessita que o classificador hoje existente no Alice para se obter as informações sobre licitações cujo tema seja tecnologia da informação tenha um menor número de falsos positivos.

5.2 Avaliação da situação

Por meio do banco de dados BD_SIASG, é possível obter as licitações realizadas e empenhadas da Administração Pública Federal direta, autárquica e fundacional. Nessa base de dados, cada licitação possui um ou mais itens, que por sua vez é associado a um elemento de despesa. O elemento de despesa tem por

finalidade identificar os objetos de gasto de cada despesa, tais como vencimentos e vantagens fixas, juros, diárias, material de consumo, serviços de terceiros prestados sob qualquer forma, subvenções sociais, obras e instalações, equipamentos e material permanentes, auxílios, amortização e outros que a administração pública utiliza para a consecução de seus fins. Quando da publicação de uma licitação, no entanto, não há, ainda, a informação relativa ao elemento de despesa, uma vez que ainda não houve empenho.

A obtenção das licitações sobre Tecnologia da Informação será realizada com base no objeto de cada licitação. No entanto, uma mesma licitação pode abordar diferentes temas, como, por exemplo, TI e obras.

Além da restrição sobre o tema de cada licitação, Tecnologia da Informação para o Siasg e para a Sefti são conceitos diferentes, uma vez que a Sefti não é a única unidade do Tribunal responsável pela fiscalização desse tema. A Secretaria de Controle Externo de Aquisições Logísticas, por exemplo, possui como atribuição exercer o controle por meio dos "processos que tratem de licitações e contratos da área-meio, cuja responsabilidade seja de órgão ou entidade com atuação em âmbito nacional e sede em Brasília. Dessa forma, a compra de equipamentos, como, por exemplo, teclados e monitores, é avaliada por esta secretaria e não pela Sefti.

5.3 Objetivos da Mineração de Dados

Considerando o objetivo do negócio de reduzir o número de falsos positivos, o objetivo da Mineração de Dados é identificar os editais de licitação sobre Tecnologia da Informação com uma precisão maior do que a existente no classificador utilizado no Alice, ou seja, diminuir o número de falsos positivos, sem, no entanto, aumentar o número de falsos negativos (recall).

5.4 Plano de Projeto

As atividades do plano de projeto são: i) estudo e estruturação das fontes de dados; ii) aplicação de técnicas de classificação textual nos dados captados; iii) avaliação dos resultados obtidos após a classificação textual.

6. Entendimento dos dados

A área de compras governamentais está organizada na forma de sistema, integrado por unidades administrativas distribuídas por todos os ministérios, autarquias e fundações públicas da administração federal. Trata-se do Sistema de Serviços Gerais – SISG, cujo órgão central é a Secretaria de Gestão da Secretaria Especial de Desburocratização, Gestão e Governo Digital – SGD, que compõe a estrutura do Ministério da Economia - ME. O SISG abrange diversos ministérios, as Secretarias da Presidência da República e mais de 300 autarquias e fundações públicas. Esse Sistema organiza a gestão das atividades de serviços gerais, o que inclui as licitações, contratações, transportes, comunicações administrativas, documentação e administração de edifícios públicos e de imóveis funcionais (SLTI/MPOG, 2002).

O Sistema Integrado de Administração de Serviços Gerais – SIASG, instituído pelo art. 7º do Decreto 1.094, de 23 de março de 1994, é o sistema informatizado de apoio às atividades operacionais do SISG, cuja finalidade é integrar os órgãos da Administração Pública Federal direta, autárquica e fundacional, e onde são realizadas as operações das compras governamentais. Ademais, os serviços de operação do Siasg são prestados pela empresa pública Serviço Federal de Processamento de Dados – Serpro (Portal do Siasg, 2020).

A criação do Siasg, em 1994, atendia à necessidade de informatizar as rotinas de serviços gerais e ocorreu simultaneamente à instituição do SISG. Dessa forma, o Governo Federal dotava a área de serviços gerais de uma estrutura organizacional uniforme, sob a coordenação de um órgão central responsável pela normatização e supervisão técnica. Ao mesmo tempo, introduziu a ferramenta informatizada como instrumento da modernização dessa atividade. Os esforços iniciais estiveram direcionados para a constituição do catálogo de materiais e serviços, o cadastramento unificado de fornecedores e o registro de preços de bens e serviços. Estas três prioridades deram origem aos três módulos inicialmente implantados: o Catálogo Unificado de Materiais e Serviços - Catmat/Catser, o Sistema de Cadastramento Unificado de Fornecedores - SICAF e o Sistema de Registro de Preços - SIREP (SLTI/MPOG, 2002).

O Siasg é constituído por módulos que realizam um conjunto de procedimentos do processo de compras e contratações compreendendo: o cadastro de fornecedores, o catálogo de materiais e serviços, o sistema de divulgação eletrônica de licitações, o sistema de registro de preços praticados, o sistema de gestão de contratos, o sistema de emissão de ordem de pagamento (Empenho), o pregão eletrônico e a cotação eletrônica, uma ferramenta de comunicação entre os seus usuários e um extrator de dados estatísticos (Datawarehouse). Os módulos estão conectados à plataforma web, dispondo de aplicativos para acesso por meio da Internet, tendo como ponto de entrada o portal Comprasnet. O portal dispõe de uma base de legislação, normas e manuais sobre compras e de um aplicativo de apoio aos pregões presenciais (SLTI/MPOG, 2002). A estrutura do Siasg pode ser definida conforme a figura abaixo:

Figura 1: Estrutura do Siasg

Os órgãos e entidades que não integram o Sisg podem utilizar o Siasg, integralmente ou em módulos específicos, por meio de adesão formal para uso do sistema, mediante assinatura de termo de adesão.

Por sua vez, o Comprasnet é o portal de compras do Governo Federal na internet que permite a realização de pregões eletrônicos, a obtenção de informações sobre licitações e contratos realizados pela APF, tais como a legislação vigente e os editais publicados, além da opção para o cadastramento dos fornecedores no Sicaf, onde é feita a verificação da regularidade fiscal das empresas licitantes, entre outras coisas.

O portal oferece ainda, como informações de acesso livre ao público em geral, consultas a avisos e resultados de licitações, a extratos de contratos celebrados, a informações sobre o fornecimento de materiais e serviços para a Administração, além da possibilidade de acompanhamento de pregões e cotações eletrônicas cadastrados no sistema.

Como informações franqueadas aos fornecedores, embora abertas à consulta pública, oferece cadastramento prévio, consulta ao registro cadastral, simulador de pregão e conexão com os portais dos órgãos de arrecadação tributária. Dessa forma, o Comprasnet funciona como ferramenta de apoio informatizado para a realização de cotações eletrônicas, pregões presenciais e eletrônicos e intenções de registro de preços.

6.1 Coleta inicial dos dados

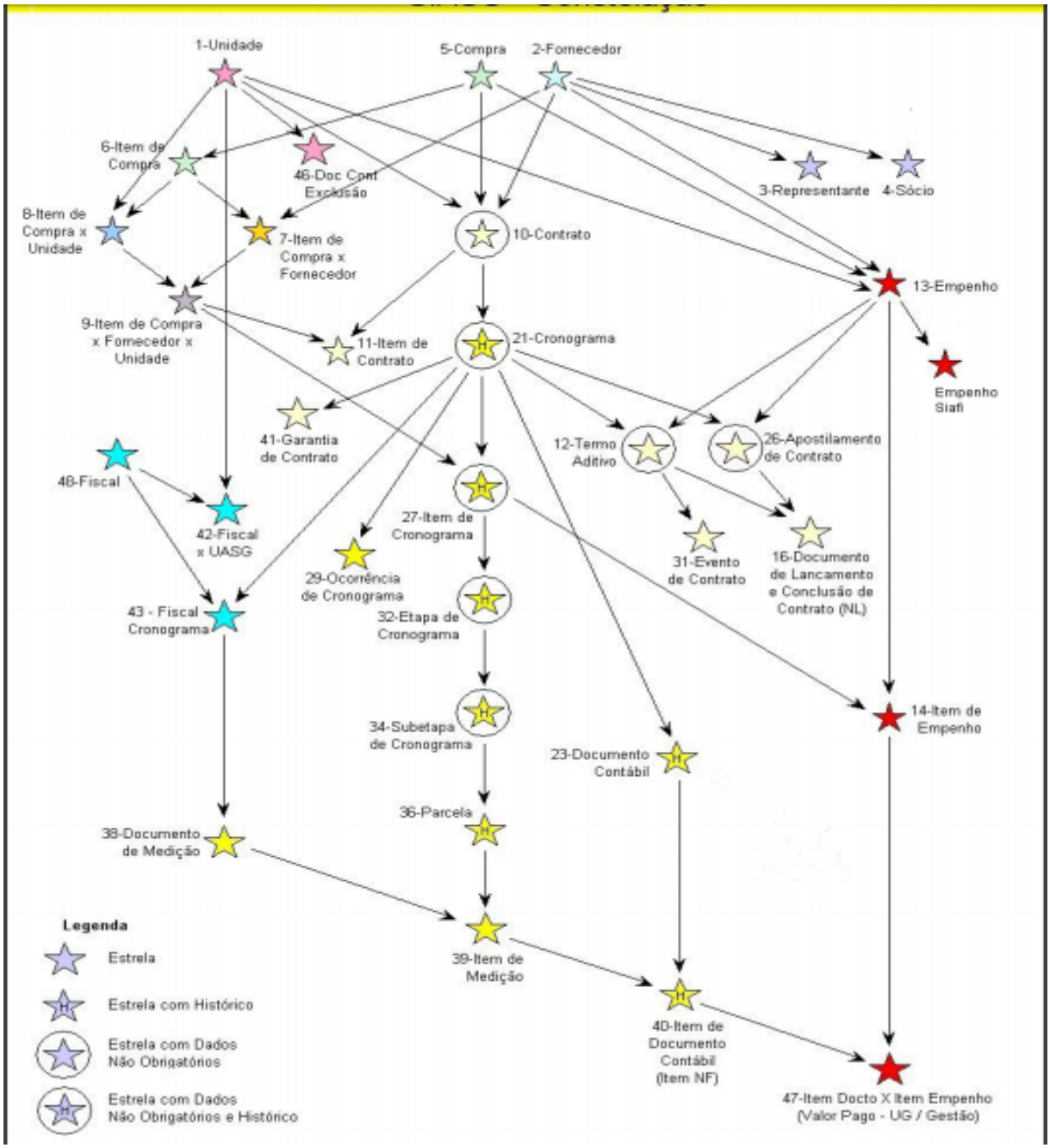
O Tribunal de Contas da União possui uma cópia do banco de dados do sistema Siasg, o DW - Compras, que é gerido pela SGD do Ministério da Economia, sua cópia é fornecida pelo SERPRO por meio de dump de banco de dados ORACLE (Via Qware) e está disponível no LabContas, com periodicidade de atualização mensal, com o nome BD_SIASG.

O TCU também possui uma cópia da base de dados do Comprasnet, o BD_COMPRASNET, que é um banco de dados transacional, gerido pelo Ministério da Economia. A cópia desse banco está disponibilizada no LabContas, com periodicidade de atualização semanal.

6.2 Descrição, exploração e qualidade dos dados

A figura 2 representa a constelação do O DW - Compras. As entidades relevantes para o projeto são: i) Compra; ii) Empenho; e iii) Item de Empenho.

Figura 2: Constelação do DW - Compras



Nessa base há atributos de diversos tipos. No entanto, os dados que serão úteis neste projeto são do tipo texto, numérico, decimal e data.

Em relação aos dados necessários para a execução do projeto, quais sejam, o objeto de uma compra e a natureza de despesa de um item de empenho, considera-se a base como de boa qualidade, uma vez que se verificou facilidade na manipulação dos dados e não foram encontrados registros com falta de dados cruciais.

7. Preparação dos dados

7.1 Seleção dos dados

A seleção dos dados foi baseada nos objetos de licitação cujos itens de empenho possuem natureza de despesa relacionada ao tema Tecnologia da Informação. O quadro abaixo descreve as tabelas do DW úteis para a seleção dos objetos.

Tabela 1: Tabelas do DW - Compras

TABELA	**Conteúdo**
D_CMPR_COMPRA	Descrição dos dados da compra como: modalidade da compra, data de ocorrência e identificação do órgão licitante.
D_EMPN_EMPENHO	Descrição dos dados de empenho, como: fonte do recurso, natureza da despesa.
D_ITEP_ITEM_EMPENHO	Descrição dos dados de item de empenho como: código do subitem, data de alteração e inclusão.
F_ITEM_EMPENHO	Relação dos dados do empenho realizado pela compra como: quantidade, compra relacionada, fornecedor.

Da mesma forma, no quadro a seguir são descritas as tabelas do Comprasnet úteis ao projeto:

Tabela 2: Tabelas do Comprasnet

TABELA	**Conteúdo**
Tbl_Pregao	Descrição dos dados de um pregão como descrição do objeto de um pregão, código da UASG responsável pelo pregão, status de um determinado pregão, etc.
Tbl_PregaoItem	Descrição dos dados de um item de pregão como código do pregão que o item está associado, quantidade de

	material ou serviço que será comprada neste item, status que se encontra determinado item de pregão.
Tbb_StatusPregaoItem	Descrição o status do item de pregão.

7.2 Limpeza, construção, integração e formatação dos dados

Foram selecionadas do BD_SIASG apenas duas colunas, a descrição do objeto de uma compra, disponível na tabela D_CMPR_COMPRA, e a coluna edital_ti, construída com base na natureza de despesa da compra analisada e com dois valores possíveis, um e zero, representando, respectivamente, TI e “Não TI”.

8. Modelagem

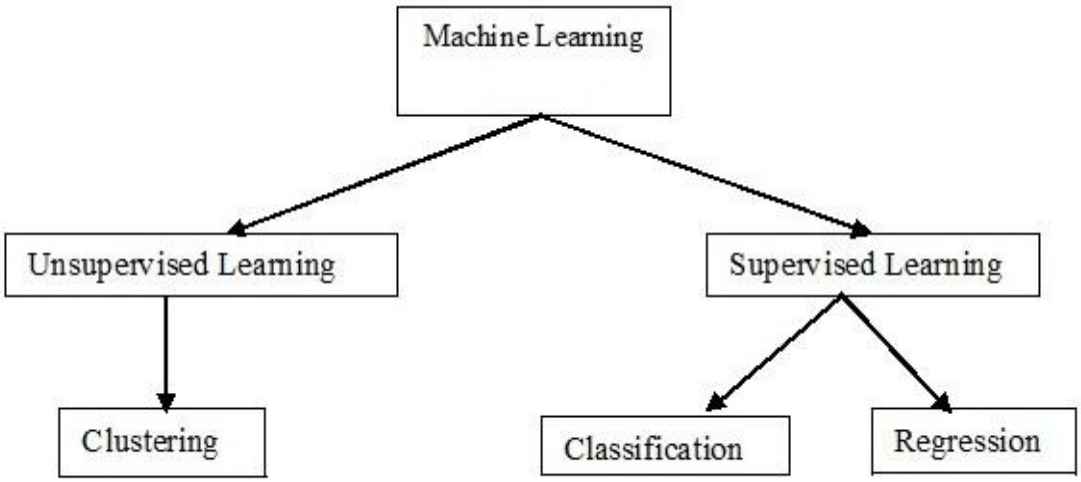
8.1 Seleção da técnica de modelagem

Aprendizagem de máquina se refere à arte de desenvolver modelos com comportamento preditivo, que são treinados através da observação de padrões em alguma fonte de dados. Um modelo está treinado quando um algoritmo de aprendizagem é utilizado em um conjunto de dados de treino com o objetivo de realizar previsões sobre novos conjuntos de dados (Helén; Persson, 2017).

Aprendizagem de máquina é categorizada em dois tipos: supervisionado e não supervisionado. Na aprendizagem supervisionada, o modelo é construído, ou seja, aprende a estimar, através dos dados de treinamento. Já na aprendizagem não supervisionada, o modelo é construído a partir de dados "não rotulados", ou seja, realiza a predição sobre novos conjuntos de dados sem nenhum conhecimento prévio de dados de treinamento (Reddy; Babu, 2018).

A aprendizagem supervisionada geralmente é realizada no contexto da classificação, quando queremos mapear a entrada para os rótulos de saída, ou regressão, quando queremos mapear a entrada para uma saída contínua. As tarefas mais comuns na aprendizagem não supervisionada são agrupamentos, aprendizado de representação e estimativa de densidade. Em todos esses casos, desejamos aprender a estrutura inerente de nossos dados sem usar rótulos fornecidos explicitamente (Soni, 2020).

Figura 3: Categorização da aprendizagem de máquina



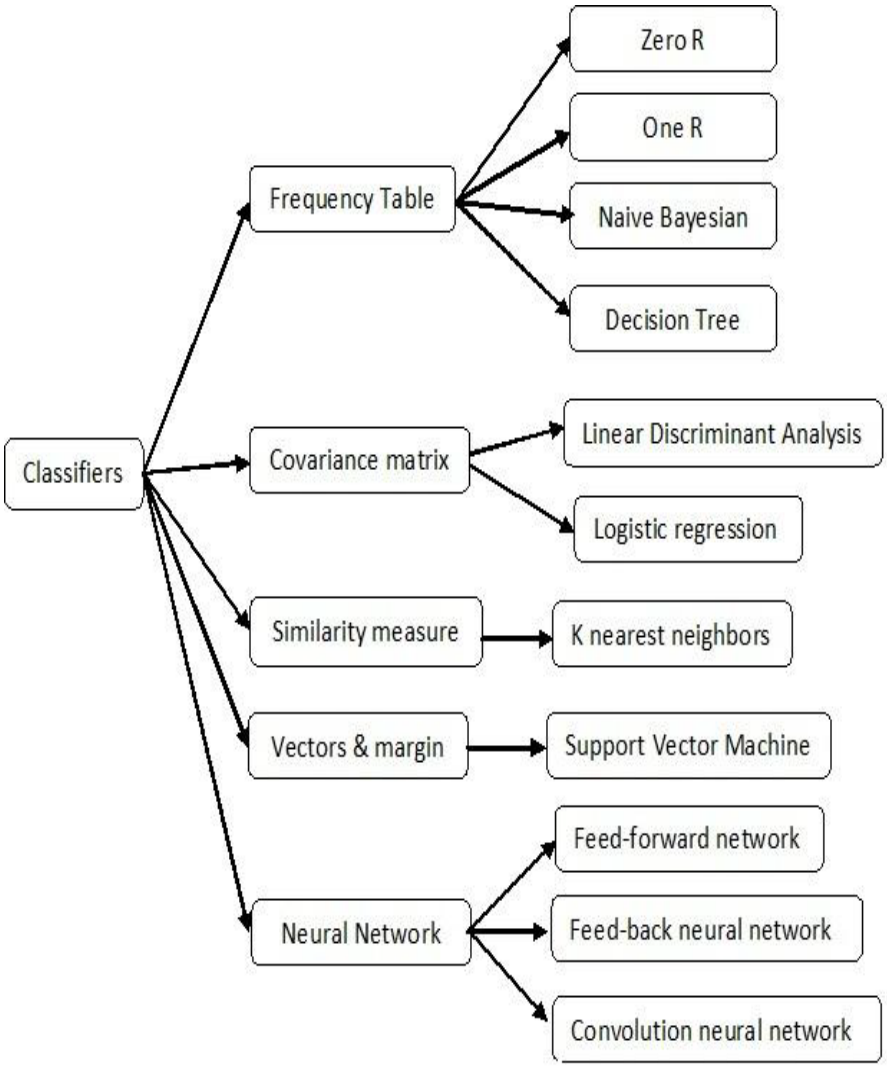
O objetivo da classificação supervisionada é induzir uma função (classificador) que seja capaz de mapear textos rotulados para seus respectivos rótulos. O classificador relaciona os termos e suas frequências com cada uma das classes. Dessa forma, o classificador é então utilizado para predizer a classe ou rotular novos textos. O modelo de classificação deve considerar o conjunto de treinamento para extrair os padrões das classes, de forma que estes sejam capazes de obter performances de classificação satisfatórias na classificação de novos textos. Para isso é necessário que os textos sejam representativos em relação às classes, de forma que o modelo de classificação seja capaz de generalizar os conceitos de classe e classificar corretamente novos exemplos (Rossi, 2015).

Na aprendizagem supervisionada, dividimos o conjunto de dados inteiro em duas partes, uma para treinamento, onde o classificador aprende com esses dados e os dados restantes são usados para testar a precisão do classificador. Feito isso, podemos testar novos dados para predizer as informações futuras desses classificadores de aprendizagem supervisionada (Reddy; Babu, 2018).

Neste projeto, uma vez que se tem acesso a conjuntos de dados já rotulados, ou seja, no BD_SIASG, é possível extrair, para cada licitação, uma relação entre o objeto pretendido e a natureza de despesa, será utilizada a aprendizagem supervisionada. Ademais, será utilizado a classificação textual para criar um classificador que seja capaz de mapear os objetos das licitações para seus respectivos rótulos, no caso deste projeto, TI ou “Não TI”.

Na literatura sobre classificação textual, os principais classificadores usados no processo de classificação de texto são o modelo Naïve Bayes, Máquinas de Vetores de Suporte (SVM), Árvores de Decisão, K Vizinhos Mais Próximos (KNN) e Redes Neurais (Kava; Desai, 2015).

Figura 4: Tipos de classificadores



Neste trabalho, decidiu-se avaliar os seguintes modelos: Random Forest, Logistic Regression, Naive Bayes, Support Vector Machines, KNN e Decision Tree. Também foram utilizados alguns metamodelos, como, por exemplo, os Ensembles.

Ademais, foi utilizado o Python como ferramenta para auxiliar a análise de dados. Uma das bibliotecas mais utilizadas do Python neste projeto foi a Scikit-Learn, que oferece um grande conjunto de modelos de classificação textual.

8.2 Geração de teste de desempenho

Foram selecionados inicialmente dez milhões de registros do BD_SIASG contendo duas colunas, objeto e rótulo (Sim ou Não). Esse conjunto de dados foi dividido em dados de treino e dados de teste na proporção de 80/20. Ou seja, 20% foram utilizados para validar o modelo. Por razões de desempenho, alguns cenários foram realizados com um número menor de dados, como, por exemplo, cinco milhões de registro, bem como com apenas um milhão.

A eficácia de uma classificação de texto pode ser avaliada em termos de precisão (p), recall (r) e medida F. Precisão é a porcentagem de documentos classificados corretamente entre todos os documentos que foram classificados como positivo pelo classificador, ou seja, medida da exatidão. Recall é definido como a porcentagem de documentos classificados corretamente entre todos os documentos pertencentes a essa categoria, ou seja, medida de completude. A medida f combina as duas medidas de recall e precisão (Kava; Desai, 2015).

Ou seja, recall é igual a seguinte fórmula: True Positives / (True Positives + False Negatives). Precisão é igual a True Positives / (True Positives + False Positives). Quanto à medida F1, é definida como 2 * (precisão * recall) / (precisão + recall).

Assim, foram utilizadas essas métricas para a seleção do algoritmo a ser utilizado na construção do modelo.

8.3 Construção do modelo

O primeiro passo da construção do modelo foi a realização da etapa de pré-processamento dos dados, fase essa que inclui a eliminação de stop words e a stemização.

Em qualquer idioma, há muitas palavras que transmitem pouco ou nenhum significado, mas são exigidos pela estrutura gramatical da língua; essas palavras são chamadas de "Stop words". Por isso, é uma prática comum a exclusão de stop words do vetor de atributos (Mahinovs; Tiwari, 2007). Dessa forma, palavras como conectores, preposições, artigos, pronomes e palavras comuns de grande ocorrência e com pouco valor para a classificação, constituem uma lista dos tokens com pouca relevância que variam de acordo com a língua do texto e serve como

parâmetro para a formação do dicionário de palavras definitivas para as demais etapas (KULTZAK, 2016).

Stemming é o processo de remover o afixo, o prefixo e o sufixo das palavras e as converter em palavras raiz (Kava; Desai, 2015). Assim, palavras que possuem variações pequenas como tempo verbal ou mudanças entre plural e singular passam por uma transformação para que seja reduzida a ocorrência de tokens com sentidos similares aumentando a eficiência durante a atribuição dos pesos para os termos (KULTZAK, 2016).

O segundo passo foi a etapa de seleção de atributos e a consequente criação do vetor de atributos, utilizando a técnica Bag of Words.

As técnicas de seleção de atributos funcionam como complemento às técnicas de pré-processamento para atenuar a dimensionalidade das representações, além de eliminar termos ruidosos e aumentar a velocidade de processamento por parte dos algoritmos. O objetivo é selecionar um subconjunto de termos de acordo com algum critério de avaliação (Rossi, 2015).

Além da seleção de um subconjunto de termos, ainda é necessário converter o texto não numérico em atributos numéricos, uma vez que a maioria dos algoritmos espera como entrada vetores de atributos numéricos com tamanho fixo, em vez de texto bruto com comprimento variável.

Esse processo é conhecido como vetorização de texto, e existem muitas maneiras diferentes de o fazer, desde a abordagem simples da vetorização baseada em contagem até técnicas mais sofisticadas como TF–IDF, Term Frequency - Inverse Document Frequency (Patel, 2020).

A representação bag of words tem como características principais a alta dimensionalidade (grande número de palavras diferentes contidas em uma coleção de texto) e a alta esparsidade (pelo fato de que grande parte das palavras ocorre somente em uma pequena parte dos documentos). Ademais, não são representadas relações entre termos ou entre documentos. Para representar relações entre termos, pode-se adotar duas estratégias: considerar frases como termos e considerar conjunto de palavras como termos. Ambas podem ser utilizadas em conjunto com bag-of-word (Rossi, 2015).

Segundo a documentação do Scikit-Learn, a representação Bag of Words da suíte scikit-learn fornece as seguintes funcionalidades para extrair recursos

numéricos de textos: tokenização, usando espaços em branco e pontuação como separadores de tokens; contador de ocorrência dos tokens em cada texto; e normalização e ponderação dos tokens em ordem decrescente de ocorrência nos textos.

A classe CountVectorizer contida na suíte Scikit-Learn, que implementa a tokenização e a contagem de ocorrência, foi utilizada neste projeto, bem como a classe TfidfTransformer.

Outra forma de extrair os atributos é utilizando TF-IDF. Term Frequency é o número de vezes que a palavra ocorre no texto. Document Frequency é o número de diferentes textos no qual a palavra ocorre pelo menos uma vez. Inverse Document Frequency é a razão entre o número total de textos e o número de textos em que o termo ocorre (Kava; Desai, 2015).

De acordo com a documentação do Scikit-Learn, em textos muito grandes, algumas palavras estarão muito presentes, portanto, carregando muito pouca informação significativa sobre o conteúdo real do texto. Dessa forma, a utilização de contagem de palavras contendo esses termos muito frequentes ocultaria a importância de termos mais raros e mais interessantes para a classificação. Dessa forma, TF-IDF é utilizado para ponderar os termos, dando mais ou menos importância a um termo presente em um determinado texto. No scikit-learn, essa normalização é implementada pela classe TfidfTransformer.

O terceiro passo foi a realização do treinamento do modelo. Para esta etapa se utilizou um subconjunto de 500.000 registros do Siasg, sendo 80% utilizados no treinamento e 20% na validação. Ademais, utilizou-se oito algoritmos de classificação textual, além de três métodos Ensemble, quais sejam:

1. SVM (Suport vector machines): SVM são máquinas de aprendizagem que podem ser consideradas fundamentadas na Teoria de Aprendizagem Estatística e utilizam em sua formulação o Princípio de Minimização do Risco Estrutural. Seu treinamento é realizado por intermédio da resolução de um QP (quadratic programming), que possui um custo computacional elevado. A principal característica das SVM's é a determinação automática dos dados de treinamento mais relevantes para o problema abordado, chamados vetores de suporte (Silva; Vieira, 2007).

2. Árvores de decisão: uma árvore de decisão é construída a partir da decomposição hierárquica do espaço dos dados gerais ou de treinamento onde uma condição no valor dos atributos é utilizada para a divisão desse espaço de forma hierárquica. Quando utilizada em textos, essa condição geralmente é representada pela presença ou falta de um ou mais termos. A divisão é feita de forma recursiva até que a árvore atinja uma quantidade mínima de folhas ou se chegue a uma condição estabelecida para a pureza da classe. Em determinado texto de treinamento a sequência de classes possíveis são aplicadas na estrutura da árvore criada a partir do seu topo até que se chegue à folha mais relevante correspondente à classe (KULTZAK, 2016).

3. K-Nearest Neighbor: a ideia deste classificador é muito simples - no espaço multidimensional, encontramos o ponto que representa o texto sendo classificado e olhamos em volta para descobrir quais outros pontos estão próximos. Apenas o número k de vizinhos mais próximos é considerado. Se todos pertencerem à mesma categoria, o novo texto também será categorizado para essa categoria. Caso contrário, a distribuição de categorias dos vizinhos mais próximos determina a probabilidade do texto pertencer àquela categoria. Em outras palavras, se, dos 5 vizinhos mais próximos, 4 pertencem à classe A e 1 pertence à classe B, o novo texto é classificado na classe A com 80% de certeza (Mahinovs; Tiwari, 2007).

4. Naive Bayes: Os classificadores Naive Bayes pertencem ao grupo de classificadores probabilísticos, que produzem uma distribuição de probabilidade sobre as classes, em vez de simplesmente exibir a qual classe um documento pertence (Helén; Persson, 2017). Os dois modelos mais utilizados são o modelo Bernoulli e o Multinomial. Ambos calculam a probabilidade de uma classe com base na distribuição das palavras no texto. Esses modelos ignoram a posição real das palavras no texto e trabalham com Bag of Words. A principal diferença entre esses dois modelos é o fato de levar ou não em consideração a frequência de palavras (Zhai; Aggarwal, 2012).

5. Logistic regression: na regressão logística, é construído um modelo de regressão para prever a probabilidade de um determinado texto pertencer à categoria numerada como "1". Assim como a regressão linear assume que os dados seguem uma função linear, a regressão logística modela os dados usando a função sigmóide, ou seja, as regressões logísticas produzem uma curva logística, que é limitada a valores entre 0 e 1. A regressão logística é semelhante a uma regressão linear, mas a curva é construída usando o logaritmo natural e não a probabilidade (Reddy; Babu, 2018).

6. Random Forest: como o próprio nome indica, consiste em um grande número de árvores de decisão individuais que funcionam como um Ensemble. Cada

árvore individual nesse modelo exibe uma classificação e a classe com mais votos se torna a previsão do modelo (Yiu, 2020).

7. **Meta-algoritmos (Ensemble):** os meta-algoritmos desempenham um papel importante na classificação por causa da sua capacidade de aumentar a precisão da classificação combinando algoritmos ou fazendo uma alteração em diferentes algoritmos. Exemplos de meta-algoritmos incluem Bagging, Stacking and Boosting (Kowsari, 2020). Alguns desses métodos alteram a distribuição dos dados de treinamento, outros combinam classificadores, e outros alteram os algoritmos com o objetivo de satisfazer critérios específicos de classificação (Zhai; Aggarwal, 2012).

Ensemble é um tipo de meta-algoritmo que i) usa diferentes subconjuntos de dados de treinamento com um único método de aprendizagem; ii) usa diferentes parâmetros de treinamento com um único método de aprendizagem; ou iii) usa diferentes métodos de aprendizagem (Kotsiantis, 2007).

O método Bagging é um método para construir ensembles que usa diferentes subconjuntos de dados de treinamento com um único classificador (Kotsiantis; Zaharakis; Pintelas, 2007). Ou seja, realiza treinamentos em vários subconjuntos aleatórios do conjunto de dados original e, em seguida, agrega suas previsões individuais (por votação ou por média) para formar uma previsão final (Dey, 2020).

Os métodos Voting e Stacking são técnicas que combinam previsões de diferentes classificadores. No método Voting, na sua forma mais simples, denominada votação por maioria, cada modelo de classificação contribui com um único voto e a previsão coletiva é decidida pela maioria dos votos, ou seja, a classe com mais votos é a previsão final. Na votação ponderada, por outro lado, os classificadores têm graus variados de influência na previsão coletiva, ou seja, a cada classificador é associado um peso específico e a previsão final é decidida somando todos os votos ponderados e escolhendo a classe com o maior valor agregado (Kotsiantis; Zaharakis; Pintelas, 2007).

O Stacking, por sua vez, utiliza a saída de vários classificadores como entrada para um outro classificador, denominado meta-classificador.

Na tabela abaixo é possível visualizar o resultado dessa etapa, qual seja, instanciação de cada um dos algoritmos listados abaixo, sem, no entanto, ajustar nenhum parâmetro dos algoritmos, e posterior treinamento utilizando os dados de treino, desbalanceados, pré-processados e transformados em vetor de atributos com Bag of Word. Após realizado o treinamento dos algoritmos, realizou-se a

validação com os dados de teste, resultando nos seguintes valores de acurácia, recall, precisão e F1:

Tabela 3: Resultado da validação dos algoritmos sem ajuste de parâmetros

Algoritmo	**Acurácia**	**Recall**	**Precisão**	**F1 Score**
LogisticRegression	0.980	0.594	0.780	0.675
ExtraTrees	0.985	0.688	0.847	0.759
RandomForest	0.985	0.700	0.846	0.766
Multinomial Naive Bayes	0.972	0.454	0.654	0.536
Bernoulli Naive Bayes	0.959	0.797	0.455	0.579
DecisionTree	0.984	0.688	0.840	0.756
KNN	0.983	0.703	0.791	0.744
SVM	0.982	0.658	0.796	0.720
Bagging (base_estimator: RandomForest)	0.985	0.694	0.842	0.761
Stacking (classifiers: [Bernoulli, ExtraTrees, KNN], meta_classifier: RandomForest)	0.985	0.699	0.841	0.763
Voting (RandomForest, Bernoulli, ExtraTrees)	0.983	0.750	0.769	0.759

Pode-se observar que os algoritmos com maior acurácia foram o ExtraTrees, o RandomForest, o Bagging e o Stacking com uma acurácia de 0,985. O Bernoulli Naive Bayes atingiu o maior recall, igual a 0,797. A maior precisão foi alcançada pelo ExtraTrees (0,847), enquanto que o maior F1 foi obtido com o RandomForest (0,766). Dessa forma, selecionou-se o algoritmo RandomForest por apresentar melhor resultado, em relação a recall e precisão.

O próximo passo da criação do modelo foi a otimização ou ajuste dos hiperparâmetros do algoritmo selecionado, por meio de GridSearch, função que permite testar combinações de parâmetros nos nossos modelos, facilitando a descoberta da melhor combinação. Um dos parâmetros dessa função é o scoring,

que define as regras de avaliação do modelo, e que pode receber como valor as strings ‘accuracy’, ‘f1’, ‘recall’, ‘precision’, etc.

Segundo a documentação do RandomForestClassifier, os principais parâmetros do RandomForest são os seguintes:

1. **n_estimators:** número de árvores na floresta
2. **criterion:** função para medir a qualidade de uma divisão (gini ou entropy)
3. **max_depth:** profundidade máxima da árvore
4. **min_samples_split:** número mínimo de amostras necessárias para dividir um nó interno
5. **min_samples_leaf:** número mínimo de amostras necessárias para estar em um nó folha
6. **max_features**: número de atributos a serem levados em consideração na procura pela melhor divisão

Após a execução da GridSearch com scoring igual a ‘f1’, verificou-se que os melhores parâmetros para o RandomForest para o conjunto de dados testado eram os seguintes: max_depth: 6, max_features: auto, min_samples_leaf: 1, min_samples_split: 0.1, n_estimators: 100. Assim, realizou-se o treinamento do modelo, dessa vez com os parâmetros descobertos com a função GridSearch, obtendo o seguinte resultado:

Tabela 4: Resultado da execução do RandomForest ajustado em relação à métrica F1

	Acurácia	**Recall**	**Precisão**	**F1**
RandomForest	0.985	0.700	0.846	0.766
RandomForest otimizado em relação a métrica F1	0.989	0.758	0.824	0.790

Tabela 5: Matriz de Confusão da execução do RandomForest em relação a métrica F1

	Classe prevista	
	Não TI	**TI**

Classe real	**Não TI**	1.585.042 (TN: Verdadeiro Negativo)	7.177 (FP: Falso Positivo
	TI	10.721 (FN: Falso Negativo)	33.628 (TP: Verdadeiro Positivo)

Após essa etapa, utilizou-se a técnica SMOTE (Synthetic Minority Over-sampling Technique) para balancear os dados. Em tarefas de classificação de dados, existem limitações que podem prejudicar o desempenho de alguns algoritmos tradicionais de aprendizagem de máquina, como, por exemplo, o desbalanceamento das amostras das classes de um conjunto de dados. Para mitigar tal problema, algumas alternativas têm sido alvos de pesquisas nos últimos anos, tal como o desenvolvimento de técnicas para o balanceamento artificial de dados, a modificação de algoritmos para que possam lidar com dados desbalanceados e a proposta de novos algoritmos para tal (Barella; Carvalho, 2015).

No conjunto de dados utilizados neste projeto, considerando quinhentos mil registros, apenas 18.033 são rotulados como TI, ou seja, apenas 3,6%. Considerando os dez milhões de registros utilizados para treinar o modelo final, 337.088 foram rotulados como sendo de tecnologia da Informação, apenas 3,37%.

Dessa forma, é possível perceber um desbalanceamento das classes (TI e “Não TI”), o que pode prejudicar o desempenho do algoritmo selecionado para a criação do modelo. Assim, utilizou-se a técnica oversampling, disponível na Imblearn, toolbox disponível para conjuntos de dados desbalanceados em aprendizagem de máquina.

Oversampling e undersampling são técnicas utilizadas em aprendizagem de máquina para ajustar a distribuição das classes em um conjunto de dados. Com undersampling, algumas amostras da classe dominante são descartadas. Por outro lado, no oversampling, alguns registros da classe minoritária são replicados.

Como resultado da execução do RandomForest sobre os dados balanceados, obteve-se o seguinte resultado:

Tabela 6: Resultado da execução do RandomForest ajustado em relação à métrica F1 e utilizando SMOTE

	Acurácia	Recall	Precisão	F1
RandomForest	0.985	0.700	0.846	0.766
RandomForest otimizado em relação a métrica F1	0.989	0.758	0.824	0.790
RandomForest otimizado em relação a métrica F1 e utilizando SMOTE	0.966	0.928	0.500	0.650

Para esse cenário, foi utilizado um conjunto menor de dados, com um milhão de registros, sendo um décimo como conjunto de teste.

Percebe-se que o recall obtido é de quase 93%, mas a precisão, no entanto, diminui para 50%. Uma vez que é preferível para a Sefti um recall maior, uma vez que não deixaria de analisar os pregões de TI, esse parece ser o melhor modelo encontrado.

A matriz de confusão retornada da execução acima foi a seguinte:

Tabela 7: Matriz de Confusão da execução do RandomForest em relação a métrica F1 e utilizando SMOTE

		Classe prevista	
		Não TI	**TI**
Classe real	**Não TI**	93.444 (TN: Verdadeiro Negativo)	3.154 (FP: Falso Positivo
	TI	244 (FN: Falso Negativo)	3.158 (TP: Verdadeiro Positivo)

Matriz de Confusão é uma ferramenta muito usada para avaliações de modelos de classificação em Aprendizado de Máquina pois permite a visualização do desempenho de um algoritmo. Da análise dessa matriz, percebe-se, por exemplo, que, dos 3.402 objetos rotulados como TI, o modelo classificou erradamente 244 objetos como não sendo de TI, mas, por outro lado, classificou corretamente 3.158, obtendo dessa forma, um recall de 92,8%.

Executou-se a função GridSearch uma segunda vez, agora com scoring igual a ‘recall’, obtendo os seguintes parâmetros: max_depth: 15, max_features: 2,

min_samples_leaf: 1, min_samples_split: 2, n_estimators: 400. Ao realizar o treinamento com esses parâmetros sobre dados balanceados, obteve-se o seguinte resultado:

Tabela 8: Resultado da execução do RandomForest ajustado em relação à métrica recall e utilizando SMOTE

	Acurácia	**Recall**	**Precisão**	**F1**
RandomForest	0.985	0.700	0.846	0.766
RandomForest otimizado em relação a métrica recall	0.370	0.898	0.046	0.088

Ou seja, os resultados não foram satisfatórios, porque as métricas pioraram em relação aos testes anteriores.

Por fim, o último passo será a utilização desse modelo treinado para realizar predições sobre as novas licitações obtidas a partir do Comprasnet, uma vez que essa base de dados possui atualização mais rápida que a base do Siasg. No início desse projeto, a atualização era semanal, mas estavam em conversas com os gestores para realizar a atualização diariamente.

8.4 Avaliação técnica

Para a seleção do melhor algoritmo a ser utilizado na criação do modelo, foram avaliadas duas métricas, recall e precisão. Quanto ao recall, apesar de possuir o melhor resultado dentre os algoritmos testados, o método Bernoulli apresentou o pior resultado no quesito precisão, apenas 0,495. Caso fosse selecionado, apresentaria uma quantidade muito grande de falsos positivos, não atingindo os objetivos do negócio.

Por outro lado, o RandomForest apresentou o melhor resultado em relação à métrica F1, que leva em consideração tanto o recall quanto a precisão. Dessa forma, esse foi o algoritmo selecionado para a criação do modelo final.

Uma vez selecionado o algoritmo, utilizou-se a função GridSearch otimizar o classificador, ou seja, realizou-se a seleção e o ajuste parâmetros do RandomForest com vistas a melhorar o desempenho atingido. Essa função foi executada em

relação às métricas recall e F1, tendo melhores resultados quando selecionada a segunda métrica.

Utilizou-se também Oversampling para balancear as duas classes existentes nos dados, TI e “NãoTI”, uma vez que a classe TI corresponde a apenas 3% de todos os registros.

9. Avaliação e trabalhos futuros

A acurácia, o recall e a precisão obtidos foram de, respectivamente, 96,6%, 92,8% e 50%. Os valores obtidos de recall estão alinhados com os objetivos de negócio, qual seja, não permitir que nenhum edital de pregão eletrônico cujo tema seja Tecnologia da Informação não seja acompanhado pela unidade técnica. No entanto, esperava-se um aumento da precisão, o que seria importante para diminuir o número de falsos positivos, o que atrapalha o acompanhamento das aquisições.

Entende-se que a rotulação dos dados pela própria Sefti irá propiciar melhores resultados uma vez que:

1. Tecnologia da Informação para o Siasg e para a Sefti, apesar de serem parecidos, são conceitos ligeiramente diferentes;
2. Alguns registros do Siasg foram rotulados incorretamente, ou seja, há casos de erro do gestor.
3. Uma vez que um mesmo objeto pode possuir vários itens, e considerando que rótulo de um objeto é derivado dos itens, dessa forma, há replicação de objetos com rótulos diferentes na base do Siasg.

Outra solução a ser considerada é a seleção de todos os registros rotulados como TI no Siasg e a concatenação desses registros com outros registros rotulados como não TI. Dessa forma, a quantidade de registros de TI seria maior, uma vez que a base é desbalanceada.

A classificação manual de registros classificados erradamente pelo classificador também é uma opção a ser testada em trabalhos futuros.

Referências bibliográficas

Rossi, Rafael Geraldeli. Classificação automática de textos por meio de aprendizado de máquinas baseado em redes. 2015.

Zhai, ChengXiang; Aggarwal, Charu C. A Survey of Text Classification Algorithms. 2012.

Silva, Cassiana Fagundes; Vieira, Renata. Categorização de Textos da Língua Portuguesa com Árvores de Decisão, SVM e Informações Lingüísticas. 2007.

Soni, Devin. Supervised vs. Unsupervised Learning. Disponível em https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8. Acesso em 5 de março de 2020.

KULTZAK, ADRIANO FRANCISCO. Categorização de textos utilizando algoritmos de aprendizagem de máquina com Weka. 2016.

Kava, Khyati S.; Desai, Nikita P. A Survey On Text Categorization Of Indian And Non-Indian Languages Using Supervised Learning Techniques. 2015.

Helén, Ludvig; Persson, Alexander. Automating Text Categorization with Machine Learning: Error responsibility routing in a multi-layer hierarchy. 2017.

Patel, Jay M. Natural language processing (NLP): text vectorization and bag of words approach. Disponível em http://jaympatel.com/2019/02/natural-language-processing-nlp-text-vectorization-and-bag-of-words-approach/. Acesso em 5 de março de 2020.

Mahinovs, Aigars; Tiwari, Ashutosh. Text Classification Method Review. 2007.

Documentação do Scikit-Learning. Disponível em https://scikit-learn.org/stable/modules/feature_extraction.html. Acesso em 5 de março de 2020.

Reddy, R. Vijaya Kumar; Babu, U. Ravi. Review on Classification Techniques in Machine Learning. 2018.

Kotsiantis, S. B.; Zaharakis, I. D.; Pintelas, P. E. Machine learning: A review of classification and combining techniques. 2007.

Yiu, Tony. Understanding Random Forest. Disponível em https://towardsdatascience.com/understanding-random-forest-58381e0602d2. Acesso em 5 de março de 2020.

Kotsiantis, S.B. Supervised Machine learning: a review of classification techniques. 2007.

Dey, Debomit. ML Bagging classifier. Disponível em https://www.geeksforgeeks.org/ml-bagging-classifier/. Acesso em 5 de março de 2020.

Kowsari, Kamran. Text Classification Algorithms: A Survey. Disponível em https://medium.com/text-classification-algorithms/text-classification-algorithms-a-survey-a215b7ab7e2d. Acesso em 5 de março de 2020.

CITADINI, Antônio Roque. O Controle Externo da Administração Pública. Ed. Max Limonad, 1995, p. 39.

Nagel, José. O Controle Externo, a estrutura e o funcionamento do TCU. 2015.

Documentação do RandomForestClassifier. Disponível em https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Acesso em 11 de março de 2020.

Barella, Victor Hugo; Carvalho, André C. P. de L. F. Tratando dados desbalanceados em classificação hierárquica. 2015.

Sistemática de análise das informações fornecidas por meio dos e-mails diários do sistema Alice para as unidades técnicas do Tribunal de Contas da União. 2018.

Chapman, P. et al. CRISP-DM 1.0: Step-by-step data mining guide. Disponível em https://www.the-modeling-agency.com/crisp-dm.pdf. 2000. Acesso em 11 de março de 2020.

Portal do Siasg. Disponível em https://www.comprasgovernamentais.gov.br/index.php/sisg/siasg. Acesso em 12 de março de 2020.

SLTI/MPOG; SIASG/Comprasnet: A Tecnologia da Informação na Gestão das Compras Governamentais na Administração Pública Federal Brasileira. 2002.