Luiz Rodrigo Airosa Castro

TÍTULO: Implementação de um Classificador de Documentos Públicos de Licitações do Portal de Compras do Governo Federal

Brasília

2020

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

LUIZ RODRIGO AIROSA CASTRO

TÍTULO: Implementação de um Classificador de Documentos Públicos de Licitações do Portal de Compras do Governo Federal

Trabalho de conclusão do curso de pós-graduação *lato sensu* em Análise de Dados para o Controle, realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista.
Orientador: Prof. Dr. Edans Flávius de Oliveira Sandes

Brasília

2020

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

REFERÊNCIA BIBLIOGRÁFICA

CASTRO, Luiz Rodrigo Airosa. **Título**: **Implementação de um Classificador de Documentos Públicos de Licitações do Portal de Compras do Governo Federal**. Trabalho de Conclusão de Curso (Especialização em Análise de Dados para o Controle) – Escola Superior do Tribunal de Contas da União, Instituto Serzedello Corrêa, Brasília, 2020.

CESSÃO DE DIREITOS

NOME DO AUTOR: Luiz Rodrigo Airosa Castro
TÍTULO: Implementação de um Classificador de Documentos Públicos de Licitações do Portal de Compras do Governo Federal
GRAU/ANO: Especialista/2019

É concedido ao Instituto Serzedelo Corrêa (ISC) permissão para reproduzir cópias deste Trabalho de Conclusão de Curso e emprestar ou vender tais cópias somente para propósitos acadêmicos e científicos. Do mesmo modo, o ISC tem permissão para divulgar este documento em biblioteca virtual, em formato que permita o acesso via redes de comunicação e a reprodução de cópias, desde que protegida a integridade do conteúdo dessas cópias e proibido o acesso a partes isoladas desse conteúdo. O autor reserva outros direitos de publicação e nenhuma parte deste documento pode ser reproduzida sem a autorização por escrito do autor.

_____

Luiz Rodrigo Airosa Castro
luiz.rodrigo@gmail.com

Ficha catalográfica

A ficha de identificação é elaborada pelo próprio autor.

Orientações em:

https://portal.tcu.gov.br/biblioteca-ministro-ruben-rosa/servicos/normalizacao-de-publicacoes.htm

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

LUIZ RODRIGO AIROSA CASTRO

TÍTULO: Implementação de um Classificador de Documentos Públicos de Licitações do Portal de Compras do Governo Federal

Trabalho de conclusão do curso de pós-graduação lato sensu em Especialização em Análise de Dados para o Controle, realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista.

Brasília, 19 de março de 2020.

Banca Examinadora:

_____

Prof. Edans Flávius de Oliveira Sandes, Dr.

Orientador

Escola Superior do Tribunal de Contas da União

_____

Prof. Saul Campos Berardo, Me.

Escola Superior do Tribunal de Contas da União

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

AGRADECIMENTOS

Em primeiro lugar, a Deus por todas as oportunidades recebidas e bençãos obtidas em minha vida.

Em especial à minha esposa Crislany, não somente pela compreensão em todas as vezes que não pude lhe dar a atenção necessária, bem como por todo o apoio, incentivo e carinho nos momentos mais difíceis e cansativos!

A meus pais, Luiz Otávio e Rita, por desde criança me mostrarem a importância da educação, bem como da capacitação contínua, ao longo da minha vida profissional.

Ao ex-Secretário da Selog, Fred, por autorizar minha participação nesse curso, bem como pela confiança da atual Secretária Tânia ao me confiar processos relevantes à unidade, a despeito de ter as segundas-feiras comprometidas em virtude das aulas presenciais. Agradeço, ainda, ao Milton e à Sorhaya por toda a compreensão e parceria durante o período.

Aos colegas de turma, responsáveis por tornarem o processo todo mais leve e divertido, além de elucidar dúvidas, em especial ao amigo Daniel pela parceria em tantos trabalhos e atividades.

E, por fim, mas não menos importante, ao meu orientador Edans, por todo o auxílio, paciência e aprendizado, não somente ao longo da elaboração deste trabalho de conclusão de curso, mas, também, ao longo de todas as disciplinas ministradas.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Aprender é a única coisa de que a mente nunca se cansa, nunca tem medo e nunca se arrepende. (Leonardo da Vinci)

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

RESUMO

Alice (Sistema de Análise de Licitações e Contratos) é um sistema do Tribunal de Contas da União (TCU) que tem como funcionalidade consolidar, resumir e agrupar informações e alertas sobre certames licitatórios publicados diariamente no Portal de Compras do Governo Federal. A validade e a pertinência dos alertas estão ligadas a tipologias dependentes do contexto em que são localizadas. Embora tenha trazido inegáveis ganhos à produtividade dos auditores, algumas evoluções podem aperfeiçoar o sistema. Entre essas evoluções, destaca-se a implementação de um classificador de documentos licitatório, que auxiliará o sistema a aumentar a eficácia e a efetividade dos seus alertas. O classificador proposto usa técnicas de mineração de dados textuais, como normalização e *bag of words*. Testou-se três modelos possíveis (*Naive Bayes*, Regressão Logística e Floresta Aleatória) com diversos parâmetros distintos, de modo a identificar a solução com melhor custo-benefício. A rotulação de dados foi feita usando como critério os nomes dos arquivos constantes da base de treino. As bases de teste e treino superaram as acurácias alvo, em especial para o algoritmo Floresta Aleatória. Usou-se ainda uma base reduzida de validação, obtendo-se uma acurácia próxima daquela desejada.

Palavras-chave: Alice. Licitação. Classificador. Mineração de Dados. Floresta Aleatória.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

ABSTRACT

Alice (Sistema de Análise de Licitações e Contratos) is a computer informational system powered by the Tribunal de Contas da União (TCU), whose functions are to consolidate, sumarize, group and issue alerts on bidding procedures published daily on the Portal de Compras do Governo Federal. The validity and relevance of the alerts are related to typologies depending on the context in which they are located. Although it brought undeniable gains to the auditors' productivity, some evolutions can improve the system. Among these developments, there´s an implementation of a bidding document classifier, which will help the system to increase the effectiveness and effectiveness of its alerts. The proposed classifier uses textual data mining techniques, such as normalization and bag of words. Three possible models were tested (Naive Bayes, Logistic Regression and Random Forest) with several different parameters, in order to identify the most cost-effective solution. The data labeling was done using the names of the files in the training base as criteria. The test and training bases exceeded the target accuracy, especially the Random Forest algorithm. A reduced basis of validation was also used, obtaining an accuracy close to that desired.

Keywords: Alice. Bidding Procedure. Classifier. Data Mining. Random Forest.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

LISTA DE ILUSTRAÇÕES

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

LISTA DE TABELAS

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

LISTA DE ABREVIATURAS E SIGLAS (opcional)

AGU - Advocacia-Geral da União

Alice –Análise de Licitações e Editais

APF – Administração Pública Federal

AUFC – Auditor Federal de Controle Externo

BB – Banco do Brasil S.A.

CAPTCHA - *Completely Automated Public Turing Test to Tell Computers and Humans Apart*

CD – Câmara dos Deputados

CEF – Caixa Econômica Federal

Comprasnet – Portal de Compras do Governo Federal

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

CSV - *Comma Separated Values*

DOCX – *Office Open XML Document*

Eletronorte - Centrais Elétricas do Norte do Brasil S.A.

PDF – *Portable Document Format*

Petrobrás – Petróleo Brasileiro S.A.

RDC – Regime Diferenciado de Contratações Públicas

RTF - *Rich Text File*

Seges - Secretaria Especial de Desburocratização, Gestão e Governo Digital do Ministério da Economia

Selog – Secretaria de Controle Externo de Aquisições Logísticas

Serpro - Serviço Central de Processamento de Dados

SGI - Secretaria de Gestão de Informações para o Controle Externo

SIASG - Sistema Administrativo de Serviços Gerais

SISG – Sistema de Serviços Gerais

TCU – Tribunal de Contas da União

TXT – *Text Document*

UASG - Unidade Administrativa de Serviço Geral

UT – Unidade Técnica do Tribunal de Contas da União

UTF-8 - *8-bit Unicode Transformation Format*

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

SUMÁRIO

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

1 INTRODUÇÃO

O Portal de Compras do Governo Federal (Comprasnet) é um sistema *web* destinado à realização de licitações (abrangendo desde o aviso de que uma licitação foi publicada até sua homologação pela autoridade competente), e permitindo, ainda, o acompanhamento de execuções contratuais (informando não apenas a data de assinatura dos contratos, mas, também, eventuais apostilamentos, aditivos e até mesmo empenhos). Ele é mantido pelo Serviço Federal de Processamento de Dados - Serpro, e consolida os procedimentos licitatórios realizados pelos participantes do Sistema de Serviços Gerais - SISG, disciplinado pelo Decreto 1.094/1994, e os de órgãos e/ou entidades que mesmo não integrantes do SISG resolveram aderir ao sistema com o fito de realizar suas licitações.

Além do Comprasnet, há outros portais de licitações no âmbito da Administração Pública Federal (APF), a exemplo do Licitações-e (que, embora mantido pelo Banco do Brasil - BB, inclui licitações das três esferas da federação), do Petronect (mantido para licitações da Petróleo Brasileira S.A. – Petrobrás e suas subsidiárias) e do Portal de Compras da Caixa Econômica Federal.

Em 2016 foi implantando no Tribunal de Contas da União (TCU) o Sistema Análise de Licitações e Editais - Alice, o qual tem por função, coletar, organizar, estruturar, analisar e direcionar as informações relativas a procedimentos licitatórios publicados diariamente no Comprasnet.

De forma simplificada, funciona assim: o Alice coleta eventuais editais (e demais documentos anexos aos instrumentos convocatórios) de diversas modalidades licitatórias e informações acerca de atas de pregões eletrônicos publicados no Comprasnet. Em seguida, o Alice executa uma série de análises para encontrar indícios de irregularidades. Essas análises, chamadas de tipologias, são divididas em duas categorias: tipologias de texto e tipologias de cruzamento de dados. As tipologias de texto buscam padrões nas sentenças dos editais e de seus anexos que possam restringir a competitividade do certame ou gerar algum outro tipo de irregularidade. Já as tipologias de cruzamento de dados buscam potenciais problemas durante a sessão pública do pregão eletrônico.

Sempre que o Alice identifica algum indício de irregularidade, ele gera alertas que serão encaminhadas às Unidades Técnicas (UT). Cada UT define critérios (com base nos

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

conceitos de risco, materialidade e relevância) e responsáveis que farão periodicamente as análises das informações oriundas do Alice. Com base nas definições feitas por cada UT, o Alice envia as informações de interesse por e-mail aos destinatários de direito.

Desde 2016 o TCU vem implementando melhorias e adaptações neste sistema para melhor se adequar às suas necessidades de negócio, tendo iniciado com um projeto piloto com diversas secretarias, a exemplo da Secretaria de Controle Externo de Aquisições Logísticas - Selog. Destaca-se, ainda, o sucesso do projeto, que está sendo ampliado nacionalmente – Projeto Alice Nacional – já contando com parceria efetiva de tribunais de contas subnacionais.

Ao longo do tempo de implementação e aperfeiçoamento do Alice, observou-se que uma melhoria útil para o sistema seria a classificação dos documentos integrantes dos procedimentos licitatórios, que é a motivação para o presente trabalho.

Isso decorre de, entre outros fatores, algumas tipologias do Alice serem aplicáveis somente a determinadas espécies documentais, o que aperfeiçoará a eficiência do sistema, além de possibilitar a redução de alertas caracterizados como "falso positivos". Além disso, espera-se um ganho qualitativo nas análises a serem feitas pelo Auditor Federal de Controle Externo (AUFC) responsável pela leitura periódica dos alertas e e-mails enviados diariamente pelo Alice.

Destaca-se, ainda que outros sistemas, como o Sistema de Análise de Orçamento (SAO), também podem ser beneficiados pela existência de um sistema de classificação de documentos.

A legislação e a praxe administrativa elencam uma série de documentos aplicáveis aos procedimentos licitatórios, entre os quais destacamos:

i. edital;
ii. projeto básico e termo de referência (documento análogo ao projeto básico na modalidade licitatória pregão);
iii. minutas de contrato e de atas de registros de preços;
iv. declarações dos licitantes; e
v. especificações técnicas de produtos.

O rol de documentos acima, o qual não é exaustivo, foi extraído a partir da análise de normativos e de amostras de licitações publicadas no Comprasnet em 2018 e 2019. Os principais normativos utilizados foram:

i. Lei 8.666/1993;
ii. Lei 10.520/2002;

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

iii. Decreto 5.450/2002;
iv. Decreto 10.024/2019;
v. Lei 13.303/2016; e
vi. Instrução Normativa – SLTI/MPOG 5/2017.

1.1 OBJETIVOS

O objetivo geral do presente trabalho é desenvolver, implementar e analisar a performance, com base no melhor custo benefício, de um classificador de documentos de procedimentos licitatórios, usando técnicas de mineração de dados e algoritmos de aprendizagem supervisionada, de modo a auxiliar o corpo técnico do TCU nas análises documentais a serem feitas (em especial no âmbito do sistema Alice).

São objetivos específicos do presente trabalho: (i) coletar e tratar os dados presentes nos arquivos de texto obtidos a partir do Alice; (ii) buscar relações e conhecimentos por meio da análise dos dados analisados; (iii) classificar previamente a amostra de documentos escolhida; (iv) testar diversos modelos e técnicas de classificação por máquina a fim de buscar o melhor custo-benefício aplicável ao problema em tela; e (v) selecionar o(s) modelo(s) e a(s) técnica(s) que apresentem os melhores resultados.

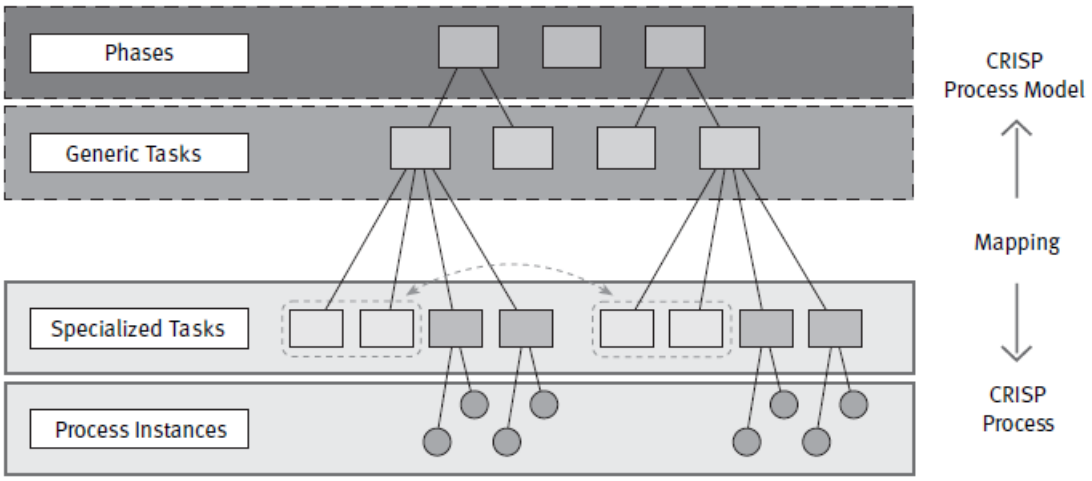
2 METODOLOGIA

Definiu-se usar como referência metodológica para o presente trabalho o *Cross-Industry Standard Process for Data Mining* (CRISP-DM), que foi desenvolvido com o intuito de funcionar como um processo padronizado, não proprietário e disponível gratuitamente, objetivando ser usado em uma grande variedade de áreas de negócio distintas, pois foi concebido com base em experimentos práticos de pessoas trabalhando em projetos de mineração de dados (CHAPMAN et al, 2000, p. 3-4).

O CRISP-DM é descrito como um modelo de processo hierárquico composto por atividades em quatro níveis distintos de abstração: fases, tarefas genéricas, tarefas especializadas e instâncias de processos (CHAPMAN et al, 2000, p. 8).

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Figura 1 - Níveis do CRISP-DM



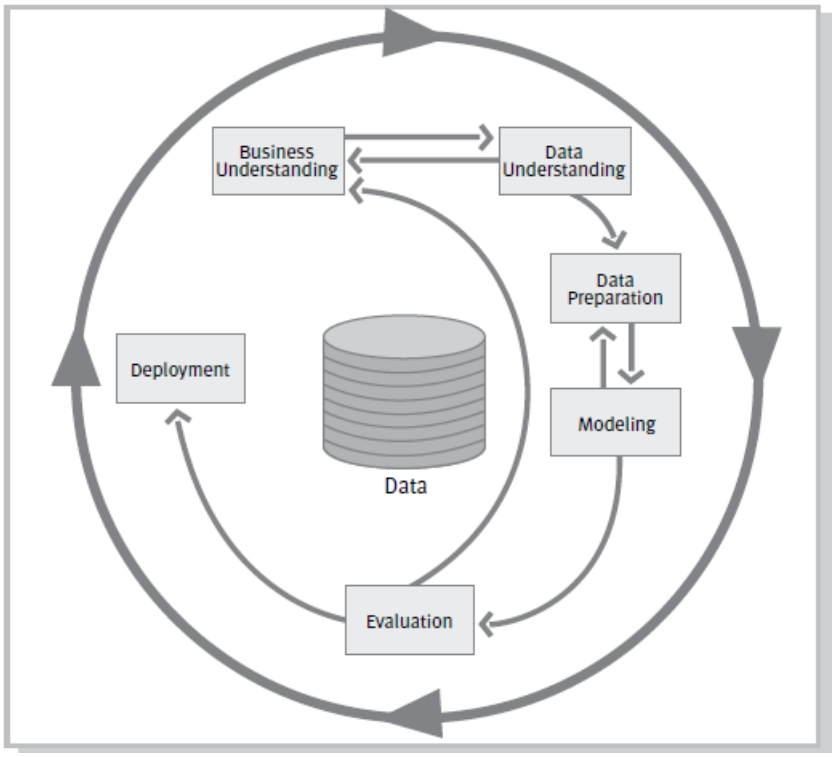
Fonte: Chapman et al (2000)

Outras metodologias similares – a exemplo do *Sample, Explore, Modify, Model, Assess* (SEMMA) – não foram selecionadas pelo fato de o CRISP-DM ser considerado mais completo, uma vez que alia tanto a construção do modelo de dados quanto a avaliação das necessidades do negócio (LAUREANO et al, 2014, p. 83-98).

O ciclo de vida dos processos no CRISP-DM é composto por seis fases, que serão explicadas a seguir, ilustradas na Figura 2.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Figura 2 - Fases do CRISP-DM



Fonte: Chapman et al (2000)

A fase de Entendimento do Negócio (*Business Understanding*) foca entender os requisitos e os objetivos a partir de uma perspectiva de negócio (MORO et al, 2011, p. 117-121).

A fase de Entendimento dos Dados (*Data Understanding*), por sua vez, trata das necessidades ao nível dos dados para o projeto, englobando atividades como: coleta inicial dos dados, descrição, exploração e avaliação de qualidade (NOGUEIRA, 2014, p. 25).

Na terceira fase, Preparação de Dados (*Data Preparation*), haverá a realização de atividades para construir o conjunto final de dados a serem modelados, o que pode englobar, entre outros, seleção de registros, formatação e conversão de dados e a criação de novos atributos (NOGUEIRA, 2014, p. 25).

A Modelagem (*Modeling*) é a fase na qual há a aplicação de diversas técnicas de modelagem existentes, bem como são calibrados os parâmetros de cada uma dessas. Nesse momento há a separação do conjunto de dados (CALIXTO et al., 2017, p. 1447-1451) em treino, a partir desse é que haverá a geração efetiva dos modelos, e teste.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Após a construção do(s) modelo(s), passa-se à fase de Avaliação (*Evaluation*), na qual deve-se analisar e revisar os passos feitos até o momento, de modo a assegurar que o modelo selecionado atenda adequadamente as perspectivas do negócio (CHAPMAN et al, 2000, p. 13).

Há, por fim, a fase de Aplicação (*Deployment*), caracterizada como aquela em que se planeja a aplicação estratégica dos resultados, bem como emprego rotineiro da mineração de dados no(s) sistema(s) de interesse (BRAZ et al, 2009, p. 1480).

Destaca-se que o processo descrito no CRISP-DM é intrinsicamente iterativo, de modo que muitas vezes uma ação em uma determinada fase acarreta mudanças ou ajustes nas demais. Por exemplo, em algumas oportunidades observou-se que a fase de Modelagem não estava apresentando resultados satisfatórios em virtude de passos insuficientes na fase de Preparação dos Dados, tendo sido necessário realizar novos processamentos a fim de obter valores mais adequados.

3 DESENVOLVIMENTO

Esta seção descreve os levantamentos de informações, decisões e procedimentos tomados desde a escolha dos dados a serem utilizados até a seleção do modelo adotado, seguindo as fases do CRISP-DM descritas no capítulo 2.

3.1 Entendimento do Negócio

Consoante o Anexo Único da Portaria – TCU 296/2018, que aprovou 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", o Alice fornece diariamente dois tipos distintos de informações às unidades técnicas do TCU, que são:

- editais de licitações (concorrência, tomada de preços, convite, Regime Diferenciado de Contratações - RDC e pregão) publicados diariamente no site do sistema Comprasnet; e
- resultados de atas de pregões divulgados diariamente no site do sistema Comprasnet.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Essas informações são disponibilizadas às UT do Tribunal por meio de dois e-mails diários. A Figura 3 ilustra um exemplo de e-mail, esse relativo aos editais publicados em uma determinada data. Em cada UT há um ou mais AUFC responsável por analisá-las.

Figura 3 - Exemplo de e-mail de licitações do Alice

6 de mar de 2020

Informe de Licitações

alice

Mostrando **17** licitações de seu interesse, de **257** publicadas em 6 de mar de 2020.

Unidade Responsável	Edital	Estimativa da Unidade*	Alertas
DEPTO. NAC. DE INFRA-ESTRUTURA DE TRANSPORTES UASG: 393003 UF: DF	Regime Dif. de Contrat. (RDC) nº63/2020 - 06/03/20 Objeto: Contratação de empresa especializada para a Elaboração de Estudos e Projetos Básico e Executivo de Engenharia para construção da Ponte Internacional Porto Xavier (Brasil) / San Xavier (Argentina), acessos - margem ...	R$ 4.593.937,53 $$$$	-
COORDENAÇÃO GERAL DE MATERIAL E PATRIMÔNIO UASG: 250110 UF: DF	Pregão nº7/2020 - 06/03/20 Objeto: Pregão Eletrônico - Locação de espaço, infraestrutura, organização, planejamento, promoção e execução de evento, gravação, compreendendo a montagem, desmontagem, limpeza, manutenção, instalações físicas, elétricas, ...	R$ 2.661.219,46 $$$	-
SENADO FEDERAL UASG: 20001 UF: DF	Pregão nº26/2020 - 06/03/20 Objeto: Pregão Eletrônico - Contratação de serviços de operação e suporte da Central de Atendimento de Telecomunicações do Senado Federal, realizados por equipe técnica residente, nas dependências da Coordenação de ...	R$ 1.353.671,04 $$$$$	-

Fonte: Elaborada pelo autor (2020)

A despeito do inegável avanço e dos ganhos de produtividade obtidos pelo TCU em decorrência das informações apresentadas pelo Alice, a Secretaria de Gestão de Informações para o Controle Externo (SGI) mapeou algumas possibilidades de melhorias para o sistema, de modo a agregar maior valor à ferramenta. Entre essas melhorias encontra-se a implementação de um módulo para classificar os arquivos coletados pelo Alice, objetivo do presente trabalho.

Esse módulo não somente auxiliará os AUFC a otimizarem a análise da documentação integrante de um certame licitatório, como o próprio sistema será aperfeiçoado. Por exemplo, vários dos alertas feitos pelo Alice versam acerca da documentação necessária para fins de habilitação técnica, que, de acordo com a jurisprudência do TCU, deve se restringir ao rol de documentos citados no art. 30 da Lei 8.666/1993. Acontece, contudo, que algumas exigências não previstas em lei podem ser exigidas em momento posterior, como no momento da assinatura do contrato. Conseguir identificar os documentos auxiliará, portanto, o Alice na

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

aplicação de suas tipologias, minimizando o risco de alertas que não se verifiquem no momento da análise do auditor competente.

O conjunto de documentos a ser classificado foi obtido por duas fontes: (i) a primeira foi a legislação correlata, cujos principais normativos estão elencados no capítulo 1 do presente trabalho; e (ii) por meio da leitura de uma série de documentos licitatórios obtidos diretamente do servidor em ambiente de produção do Alice, o qual contém documentos de textos originados das UASG e disponibilizados em outros formatos no Comprasnet. Tais documentos são detalhados no item 3.2 deste trabalho.

3.1.1 Determinando os objetivos do negócio

O objetivo primário deste trabalho é classificar de forma adequada os arquivos comumente anexados a certames licitatórios, auxiliando o Alice a aplicar com maior eficiência suas tipologias textuais, colaborando para que os AUFC tenham um maior ganho de produtividade em seus trabalhos. Isso porque empiricamente se observou um alto número de "falsos positivos" nos alertas apresentados pelo Alice, pois determinadas cláusulas são vedadas em certos momentos do procedimento licitatório, enquanto permitidas em outros, vide exemplo citado no item 3.1 do presente trabalho.

Também é um objetivo do presente trabalho possibilitar ao TCU a criação de uma base de dados com arquivos das diversas espécies de documentos existentes em processos licitatórios, uma vez que atualmente os arquivos colocados no Comprasnet pelas Unidades Administrativas de Serviços Gerais (UASG) não costumam discriminar a espécie de documento em que se enquadram.

As UASG são divisões administrativas de órgãos e entidades que integram o Sistema Administrativo de Serviços Gerais (SIASG), instituído pelo art. 7º do Decreto 1.094/1994. Na prática, as UASG são as unidades adquirentes de bens, serviços e obras licitados por meio do Comprasnet. Cada órgão – de acordo com suas particularidades – divide-se em quantas UASG considerar pertinentes.

Por exemplo, a Câmara dos Deputados (CD) possui uma única UASG (de código 010001), enquanto o TCU possui dezoito dessas unidades, vide Figura 4.

Figura 4 - Conjunto de UASG do TCU

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Cod. UASG	Nome da UASG	UF
30027	SEC.DE CONT.EXT. NO ESTADO DE RORAIMA	RR
30023	SECRET.DE CONTROLE EXTER.NO EST.DE/TO	TO
30013	SECRETARIA CONTROLE EXTERNO-TCU/ES	ES
30026	SECRETARIA DE CONTROLE EXTER.NO EST.DO ACRE	AC
30008	SECRETARIA DE CONTROLE EXTERNO EM PERNAMBUCO	PE
30010	SECRETARIA DE CONTROLE EXTERNO EM SERGIPE	SE
30011	SECRETARIA DE CONTROLE EXTERNO NA BAHIA	BA
30002	SECRETARIA DE CONTROLE EXTERNO NO PARA	PA
30024	SECRETARIA DE CONTROLE EXTERNO/TCU/AP	AP
30015	TCU-SECRETARIA CONTROLE EXTERNO/RJ	RJ
30007	TCU-SECRETARIA DE CONTROLE EXTERNO NA PARAIBA	PB
30012	TCU-SECRETARIA DE CONTROLE EXTERNO/MG	MG
30004	TCU-SECRETARIA DE CONTROLE EXTERNO/PI	PI
30019	TCU-SECRETARIA DE CONTROLE EXTERNO/RS	RS
30001	TCU-TRIBUNAL DE CONTAS DA UNIAO/DF	DF
30005	TCU/SEC. DE CONTROLE EXTERNO NO CEARA	CE
30003	TCU/SERC. DE CONTROLE EXTERNO NO MARANHAO	MA
30021	TCU_SECRETARIA DE CONTROLE EXTERNO/GO	GO

Fonte: Elaborada pelo autor (2020)

3.1.2 Avaliação da Situação

Entende-se que o TCU, em especial a SGI, possui a infraestrutura e as ferramentas necessárias para a obtenção dos dados de entrada necessários ao trabalho de classificação. Isso porque o algoritmo do Alice já trata a conversão de documentos nos mais diversos formatos (*Microsoft Word Open XML Format – DOCX, Portable Document File – PDF, Rich Text File - RTF* etc.) e os converte em arquivos do tipo *Text file – TXT*.

Além disso, a SGI armazena os documentos extraídos dos certames licitatórios, possuindo, sob sua custódia, centenas de milhares de arquivos extraídos originalmente do Comprasnet.

Ressalta-se que todos os dados extraídos o foram do Comprasnet, não incluindo certames publicados em portais similares, a exemplo do Licitações-e ou do Petronect. Isso porque cada portal possui um fluxo próprio para o *download* de documentos, bem como restrições distintas. Por exemplo,o uso de *Completely Automated Public Turing Test to Tell Computers and Humans Apart* (CAPTCHA) dificulta os procedimentos utilizados para extração de informações e arquivos, a exemplo de *web scraping* (BARREIRA, 2014, p. 3).

3.1.3 Determinando os objetivos de mineração de dados

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Será considerado como um critério de objetivo de sucesso do negócio a implementação de um classificador de documentos capaz de predizer o tipo de documento com uma acurácia superior a 90%.

Optou-se por esse valor pelo fato de se estar lidando com arquivos texto, os quais não são padronizados de forma geral, o que por si dificulta a classificação. Ainda assim, observou-se algumas iniciativas de padronização de documentos, como a elaboração de minutas de editais, termos de referência e contratos por parte da Advocacia-Geral da União (AGU). Embora essas padronizações tendam a ser isoladas em um ou poucos órgãos, eventualmente os modelos classificatórios são capazes de identificar alguns padrões de texto que caracterizam tais modelos de documento.

Vale destacar que em função dessa não padronização, foram verificados arquivos não relacionados diretamente a procedimentos licitatórios, a exemplo de organogramas de órgãos contratantes, e documentos, que embora ligados a licitações públicas, não estão previstos de forma expressa na legislação, como termos de renúncia a recursos de decisões proferidas por comissão de licitação. Esses tipos mais raros de documentos estão fora do escopo do presente trabalho.

3.2 Entendimento dos Dados

Esta fase se inicia com a coleta inicial dos dados, e abrange atividades e análises para obter familiaridade com os dados, identificados problemas de qualidade nesses, ou a detecção de subconjuntos de interesse (WIRTH, 2000, p. 29-39).

3.2.1 Coletando os Dados

A coleta dos dados foi feita mediante extração de arquivos da versão em produção do Alice no TCU. Basicamente se reuniu um conjunto de arquivos no formato texto, provenientes de 63.721 licitações, constantes da base de dados relativa aos anos de 2018 e 2019, de um total de 2.948 UASG distintas.

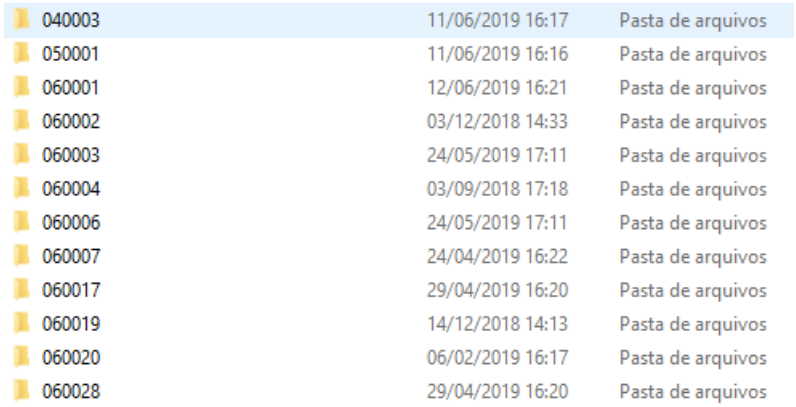
Os arquivos texto foram disponibilizados em arquivo compactado (em formato ZIP) contendo mais de 220.000 arquivos ao todo, com tamanho total de 9,01 gigabytes.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

3.2.2 Descrevendo os Dados

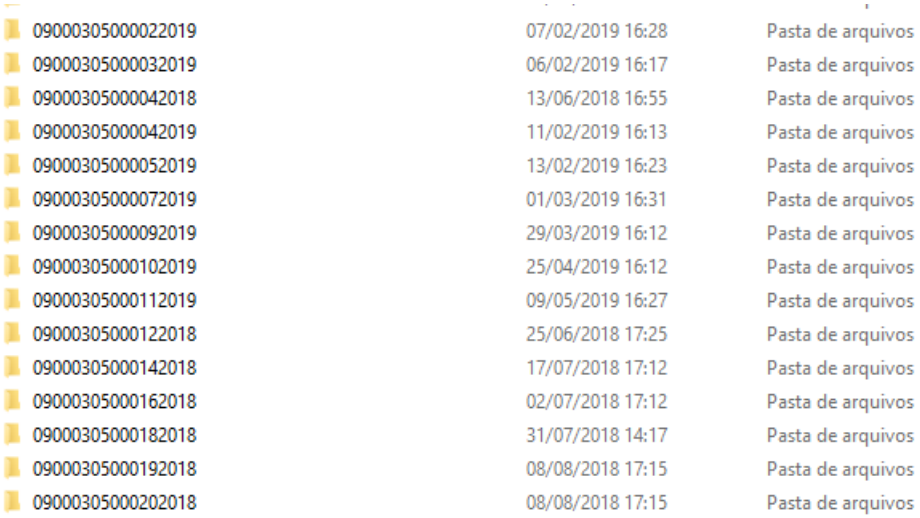
Os arquivos textos foram salvos em uma estrutura de diretórios e subdiretórios, em que cada diretório raiz correspondia a uma UASG, e cada subdiretório correspondia a uma licitação específica dessa UASG. Dentro de cada subdiretório há os arquivos coletados pelo Alice para aquela licitação, consoante as Figuras 5 a 7.

Figura 5 - Amostra do conjunto de UASG



Fonte: Elaborada pelo autor (2020)

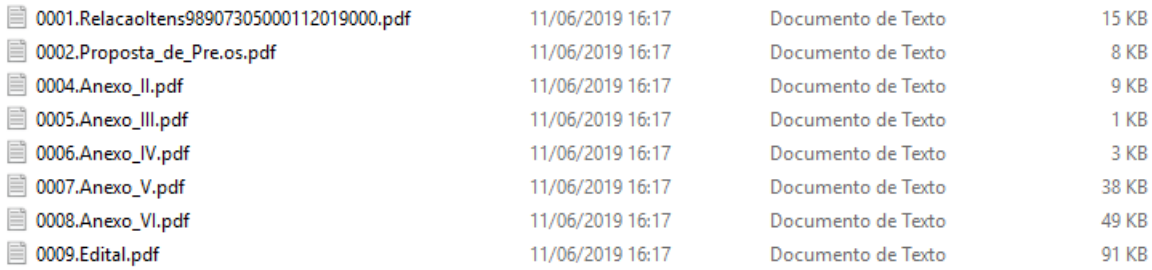
Figura 6 - Exemplos de licitações de uma mesma UASG



Fonte: Elaborada pelo autor (2020)

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

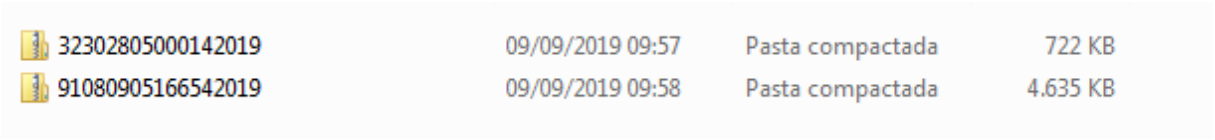
Figura 7- Conjunto de arquivos texto de uma mesma licitação



Fonte: Elaborada pelo autor (2020)

Para o entendimento adequado do trabalho, faz-se necessária uma explicação sobre como o Comprasnet opera. Cada licitação publicada no Comprasnet disponibiliza um arquivo zip contendo o edital e seus anexos.

Figura 8 - Arquivos compactados extraídos do Comprasnet



Fonte: Elaborada pelo autor (2020)

Os nomes desses arquivos compactados possuem 17 dígitos. As seis primeiras posições correspondem ao código da UASG; a sétima e a oitava posição informam a modalidade licitatória (o código '05', o mais utilizado, corresponde à modalidade pregão); as demais posições informam o número/ano da licitação.

Tabela 1 - Composição dos identificadores de licitações no Comprasnet

Posição dos Dígitos	Função
1º - 6º	Identificador da UASG contratante
7º - 8º	Modalidade licitatória do certame
9º - 17º	Identificador do número e ano da licitação

Fonte: Elaborada pelo autor (2020)

Tome-se como exemplo o certame com a identificação "03000105000012020". Os primeiros seis dígitos (030001) identificam a UASG Tribunal de Contas da União. Os dígitos sete e oito (05) fazem referência à modalidade licitatória, no caso o pregão eletrônico. Por fim,

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

os demais (000012020) especificam o número e o ano da licitação, no caso 01/2020. A título de curiosidade, a UASG que mais promoveu licitações no espaço amostral analisado (2018 e 2019) foi a de código 910809 – Centrais Elétricas do Norte do Brasil S.A. (Eletronorte) – com 2.354 licitações (3,7% do total).

Após o Alice obter os arquivos zip do Comprasnet, aquele sistema adota uma estratégia de extração dos arquivos em modo *flat*, onde não há subdiretórios nem eventuais arquivos compactados recursivos (zips dentro de zips). Assim, todos os arquivos são localizados em um mesmo diretório.

A fim de ordenar os arquivos e evitar repetições de nomes, o sistema atribui um numeral de quatro dígitos e um ponto (“.”) no início do nome de cada arquivo extraído. Assim, eles começam com sequências como “0001.”, “0002.”, “0003.” etc. Espaços e caracteres especiais nos nomes dos arquivos são convertidos para “_” e ‘ “.”, respectivamente. Por fim, arquivos de tipos como PDF, DOCX e RTF são convertidos para o formato de arquivo TXT por meio da ferramenta *Apache Tika*.

Destaca-se, ainda, que junto aos arquivos compactados, definidos por cada UASG, também é enviado um arquivo PDF, criado automaticamente pelo Comprasnet, cuja nomenclatura inicia com a *string* “RelacaoItens”.

Figura 9 - Exemplo de conteúdo de arquivo "RelacaoItens"

RELAÇÃO DE ITENS - PREGÃO ELETRÔNICO Nº 00061/2019-000 SRP

1 - Itens da Licitação

1 - Equipamento de raio x tipo industrial

Descrição Detalhada: FILMES PARA IMPRESSÃO DE RAIOS-X PROCESSAMENTO A SECO COMPATÍVEL COM PROCESSADORA AGFA – DRY STAR. TAMANHO 25X30 CM, CAIXA COM 100 UNIDADES.
Tratamento Diferenciado: Não
Aplicabilidade Decreto 7174/2010: Não
Critério de Julgamento: Menor Preço
Quantidade Total: 48
Unidade de Fornecimento: Unidade
Quantidade Máxima para Adesões: 240
Local de Entrega (Quantidade): Ji-Paraná/RO (48)

2 - Equipamento de raio x tipo industrial

Descrição Detalhada: FILMES PARA IMPRESSÃO DE RAIOS-X PROCESSAMENTO A SECO COMPATÍVEL COM PROCESSADORA AGFA– DRY STAR. TAMANHO 20X25 CM, CAIXA COM 100 UNIDADES.
Tratamento Diferenciado: Não
Aplicabilidade Decreto 7174/2010: Não
Critério de Julgamento: Menor Preço
Quantidade Total: 48
Unidade de Fornecimento: Unidade
Quantidade Máxima para Adesões: 240
Local de Entrega (Quantidade): Ji-Paraná/RO (48)

Fonte: Elaborada pelo autor (2020)

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Tal arquivo resume os itens e grupos a serem adjudicados em todas as licitações constantes no Comprasnet, independente da modalidade de licitação e/ou do total de arquivos abrangidos por cada certame. Uma vez que esse arquivo é comum a todos os certames, foi excluído das análises, posto que não colabora para a implementação do classificador de documentos almejado.

3.2.3 Explorando os Dados

Uma tarefa necessária para o problema foi responder à seguinte questão: os órgãos e entidades da APF publicam os documentos relativos a seus procedimentos licitatórios (editais, termos de referência, projetos básicos, minutas de contrato, declarações etc.) em um mesmo arquivo ou em um conjunto deles?

Para responder ao questionamento supra, usou-se um *notebook Jupyter,* utilizado a partir do *software open-source Anaconda Distribution,* para listar o total de arquivos obtidos pelo Alice por licitação, e foram obtidas as seguintes informações abaixo:

Tabela 2 - Licitações analisadas, por quantidade de arquivos

Total de Licitações	Licitações com um Único Arquivo	Licitações com Vários Arquivos
63.721	46.783	16.938

Fonte: Elaborada pelo autor (2020)

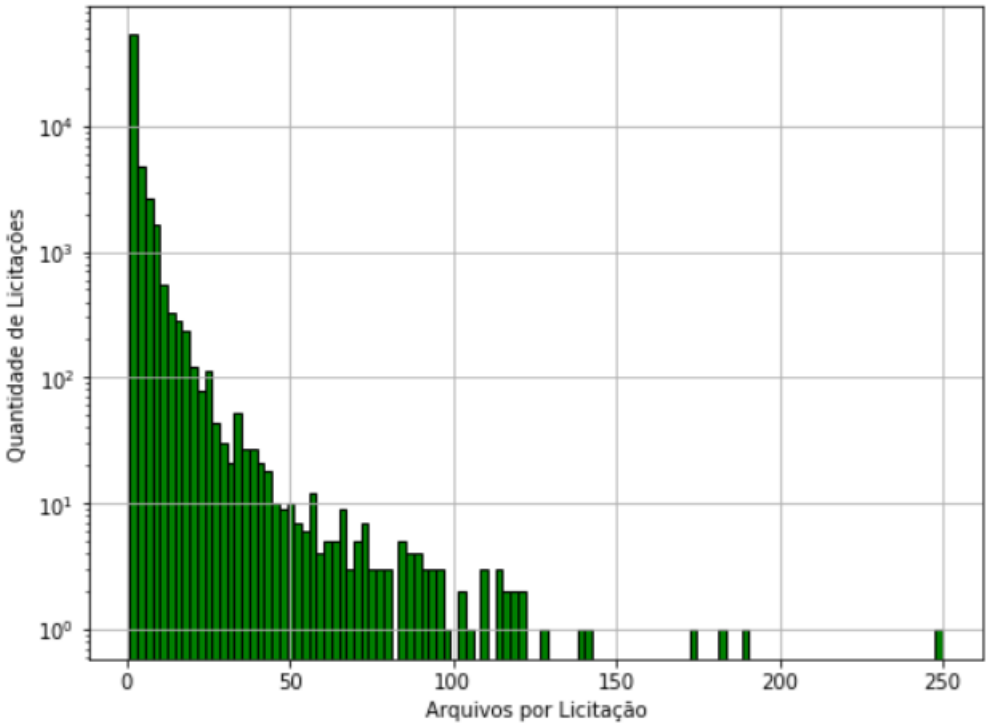
Dos dados acima, observa-se que 73,42% das licitações foram condensadas em um único arquivo, enquanto o restante (26,58%) era composto de ao menos dois arquivos.

No primeiro momento, foram excluídas das análises aqueles certames que condensam todos os documentos licitatórios em um único arquivo, porquanto eles podiam causar divergências na classificação implementada. A evolução do classificador para uma versão que analisa trechos de um único arquivo e classifica cada um desses de acordo com o seu conteúdo é uma melhoria recomendada como trabalho futuro, vide seção 4 deste trabalho.

O histograma da Figura 10 apresenta, em escala logarítmica, a distribuição do número de arquivos publicados por licitação, deixando claro que a maior parte dos certames não divide os distintos documentos em um único arquivo. Nota-se, ainda, que o máximo de arquivos distintos relacionados a uma única licitação foi de 250.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Figura 10 - Histograma de Licitações por arquivos



Fonte: Elaborada pelo autor (2020)

3.3Preparação dos Dados

Esta fase engloba todas as atividades necessárias à construção do conjunto final de dados formatados, a partir do grupo inicial de dados não tratados (AZEVEDO, SANTOS, 2008, p. 3). Foi composta pelas seguintes tarefas: seleção e exclusão de dados; formatação; criação de rótulos (rotulação) dos dados; e divisão das bases de treinamento e teste.

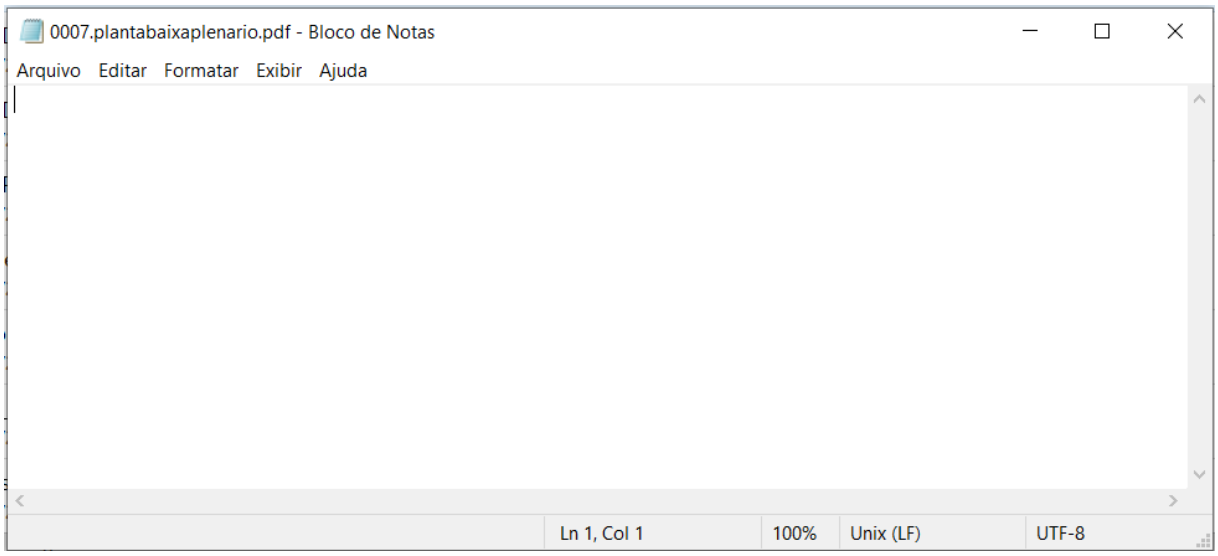
3.3.1 Seleção e Exclusão de Dados

Consoante explanado à seção 3.2.1 deste trabalho, a coleta de dados já envolveu uma pré-seleção, posto que os arquivos foram obtidos diretamente do servidor de aplicação do Alice, tendo sido feito todo o procedimento de conversão para o formato TXT, além do agrupamento de arquivos em um mesmo diretório, feito automaticamente pelo sistema.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Isso não significa, contudo, que não houve uma seleção posterior de dados a serem analisados. Observou-se, por exemplo, que arquivos que originalmente eram imagens, a exemplo de plantas de engenharia, não foram adequadamente convertidos pelo Alice, vide Figura 11.

Figura 11 - Exemplo de arquivo com conteúdo não lido pelo Alice



Fonte: Elaborada pelo autor (2020)

Assim, para fins de implementação do classificador, excluíram-se não somente as licitações que condensaram os documentos em um único arquivo, vide seção 3.2.3 deste trabalho, mas também àquelas com conteúdo inferior a cem caracteres.

O valor de cem caracteres foi definido empiricamente, após uma série de testes em busca de um valor que não fosse exageradamente restritivo, sequer permitisse a análise de documentos com conteúdo sem utilidade para os fins desejados neste trabalho. Uma sugestão de aperfeiçoamento em trabalho futuro é durante a etapa de treino permitir que os próprios modelos informem o valor mínimo de caracteres existentes nos arquivos para uma classificação com acurácia superior ou igual àquela desejada.

Inicialmente, foram selecionados os campos de interesse detalhados na tabela a seguir:

Tabela 3 - Campos de interesse para o classificador de documentos

Campo	**Descrição**
UASG	Código de seis dígitos que identifica a UASG realizadora do certame

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Campo	Descrição
ID_LICITACAO	Conjunto de 17 dígitos que identifica de forma única cada licitação
CAMINHO_ARQUIVO	Localização do arquivo na estrutura de pastas no servidor web do *Jupyter Notebook* constante do *Anaconda Navigator*. Foi necessário para acessar os conteúdos dos arquivos usados nos testes e treinos
NOME_ARQUIVO	Título do arquivo. Foi o critério adotado para rotular os dados de testes e treinos nas categorias desejados
CONTEUDO_ARQUIVO	Conjunto inicial de caracteres extraído do arquivo. É o critério usado na modelagem a fim de identificar a acurácia dos modelos testados

Fonte: Elaborada pelo autor (2020)

3.3.2 Formatando os Dados

Conforme citado à seção 3.2.2 deste trabalho, o Comprasnet usa uma string de 17 posições para identificar os procedimentos licitatórios. Uma vez que todos caracteres da *string* são números, por padrão o *Python* – e a biblioteca *Pandas* – os tratava como o tipo *INT*. Assim, foi necessário convertê-los em *string*. O mesmo se aplicando para os códigos de UASG, que são compostos por seis dígitos.

Isso foi necessário em especial por duas situações específicas, a saber: (i) existem alguns órgãos com códigos que começam com o dígito 0 – a exemplo do próprio TCU, cujo código de UASG é 030001 –; e (ii) uma vez que a quantidade de dados analisado superou o total de nove *gigabytes*, fez-se necessário o uso de arquivos *Comma Separated Values* (CSV) para a persistência e a manipulação mais eficiente dos *dataframes* usados.

Para a correta leitura dos arquivos, foi necessário usar a codificação 8-*bit Unicode Transformation Format* - UTF-8, uma vez que o *Pandas* não conseguiu ler adequadamente os

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

dados dos arquivos (coluna CONTEUDO_ARQUIVO) com o padrão de codificação padrão do sistema operacional.

Na etapa de modelagem, verificou-se a necessidade de formatar as *strings* usadas para se verificar a acurácia dos algoritmos testados. Essa formatação foi composta dos seguintes passos: (i) realizar procedimentos de normalização, lematização e *stemming*; (ii) remover *stop words*; e (iii) conversão das palavras em matrizes de números inteiros, usando o método CountVectorizer da biblioteca *scikit-learn*.

A normalização tem por intuito reduzir o conjunto de palavras existentes ao converter aquelas com significados semelhantes a um único conceito. Uma técnica comum para possibilitar isso é o *stemming*, que consiste em reduzir uma palavra a sua raiz, removendo tanto o seu prefixo quanto o seu sufixo (CHIARA, 2003, p. 146).

A lematização pode ser definida como a representação das palavras por meio de seus lemas – palavras dicionarizadas –, normalmente representado pelo infinitivo, no caso dos verbos, e pelas formas masculino singular dos substantivos e adjetivos (DE LUCCA, 2002, p. 8).

Stop words são conceituadas como palavras sem conteúdo semântico significante, não sendo relevantes para extração de informações a partir dos textos analisados (LOPES, 2004, p. 21). Comumente, classes gramaticais como preposições, pronomes e artigos são considerados *stop words*.

3.3.3 Criando dados rotulados;

Usou-se o *software Microsoft Excel* para a criação de uma nova coluna chamada CATEGORIA, a qual contém uma classificação do tipo do arquivo com base em palavras chaves encontradas no nome do arquivo. Esta coluna (*target*) será utilizada para o treinamento e testes dos algoritmos de aprendizado supervisionado. A Tabela 4 apresenta as sete classes definidas e as palavras chaves utilizadas para o processo de classificação.

Tabela 4 - Palavras-chaves para rotulação de dados

Código	Tipo de Documento	Palavras chaves
1	Edital	Edital
2	Termo de Referência / Projeto Básico	Termo_Ref / Termo_de_Ref / Projeto_B

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Código	Tipo de Documento	Palavras chaves
3	Minuta de Contrato	Contrato
4	Modelo de Proposta de Licitante	Proposta
5	Declaração de Licitante	Declar
6	Minuta de Ata de Registro de Preços	Ata_RP / Registro_de_P
7	Especificações Técnicas dos Produtos	Especifica / Modelo

Fonte: Elaborada pelo autor (2020)

Importante destacar que os sete tipos usados acima não são os únicos existentes. Ao longo da análise documental, feita por meio de amostra, foram encontrados documentos distintos, previstos ou não na legislação, como cronogramas físicos financeiros, instruções normativas, pesquisas de preços e outros (vide capítulo 3.1.3 deste trabalho). A esses outros documentos – bem como aqueles cujos nomes não foram suficientes para se atribuir um valor de categoria de 1 a 7 – foi atribuído o valor zero, e foram excluídos da etapa de modelagem.

Assim, a base final para a seleção dos conjuntos foi constituída por 54.713 arquivos, distribuídos conforme a Tabela 5.

Tabela 5 - Distribuição de arquivos por códigos de categoria

Código	Nome Categoria	Quantidade	Percentagem
1	Edital	15.398	28,14
2	Termo de Referência ou Projeto Básico	11.239	20,54
3	Minuta de Contrato	8.080	14,77
4	Modelo de Proposta de Licitante	5.674	10,37
5	Declaração de Licitante	5.950	10,87
6	Minuta de Ata de Registro de Preços	3.594	6,57
7	Especificações Técnicas de Produtos	4.778	8,73
	Total:	54.713	100,00

Fonte: Elaborada pelo autor (2020)

3.3.4 Dividindo em Conjuntos de Dados de Treinamento, Validação e Teste

A divisão dos arquivos com códigos de categoria com valores de 1 a 7 foi feita pelo método *train_test_split* do scikit-learn, usando o percentual de 75% dos dados para treinamento

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

dos modelos e 25% para a validação. Isso correspondeu, respectivamente, a 34.570 registros para treinos e 11.524 para validação. Após isso, fez-se a rotulação manual de um conjunto de documentos a fim de testar a acurácia obtida pelo modelo escolhido, vide seção 3.4.3.1 deste trabalho

Optou-se por não atribuir nenhum valor ao parâmetro *shuffle* do método *train_test_split.* Desse modo o valor *default* (*True*) foi adotado, evitando um agrupamento exagerado de dados similares (por exemplo, todas as licitações de uma mesma UASG no conjunto de treino ou no de teste), o que poderia ocasionar problemas como *underfitting* ou *overfitting*.

Ocorre o *overfitting* quando o modelo "memoriza" os padrões da base usada inicialmente, perdendo sua capacidade de generalização com novas entradas de dados. Já o *underfitting* é caracterizado pelo baixo treinamento, não permitindo ao modelo sequer aprender os padrões e comportamentos a serem generalizados (ARAÚJO, 2018, p. 150).

3.4 Modelagem

Após a obtenção e o pré-processamento de dados, procede-se à implementação de modelos para a análise e seleção da escolha pretendida à implantação do classificador desejado. A modelagem é considerada por alguns autores como a principal etapa dos processos de mineração de dados (DO NASCIMENTO, 2018, p. 6).

3.4.1 Seleção das Técnicas de Modelagem

Foram selecionados três algoritmos distintos de classificação, a fim de comparar os prós e contras de cada, bem como os resultados a serem alcançados por eles. Os algoritmos selecionados foram os seguintes: *Naive Bayes*; *Random Forest* (Floresta Aleatória); e *Logistic Regression* (Regressão Logística).

O Naive Bayes é um algoritmo que provê um classificador probabilístico, permitindo estimar a probabilidade de um elemento se enquadrar em uma determinada categoria, com base no Teorema de Bayes (OLIVEIRA, 2004, p. 317). O termo "*naive*" (ingênuo) significa que o classificador desconsidera a correlação entre as variáveis existentes. Apesar dessa simplificação, ainda assim o classificador *Naive Bayes* é considerado efetivo (DOMINGOS, 1997, p. 103-130).

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

A Floresta Aleatória é um modelo baseado em árvores de decisão, com diversas aplicações, desde diagnóstico médico por imagens (CRIMINISI, 2013, p. 1293-1303) até análise de big data (GENUER, 2017, p. 28-46). Após uma série de testes comparando a acurácia desse modelo com a acurácia dos demais, optou-se por atribuir, respectivamente, o valor de cem para o hiperparâmetro "*n_estimators*" (total de árvores usadas pelo modelo). Uma possibilidade de aperfeiçoamento do sistema, em trabalho futuro, é definir um valor otimizado para o hiperparâmetros em questão.

A árvore de decisão, das quais as florestas aleatórias são formadas, é um método adequado quando deseja-se classificar dados ou predizer valores de saída. Seu atributo mais importante é apresentado na árvore como o primeiro nó, e os menos relevantes são mostrados nos nós subsequentes. A principal vantagem de se usar árvores de decisão é tomar decisões com base nos atributos mais relevantes (LEMOS, 2005, p. 225-234).

O algoritmo de Regressão Logística funciona por meio da medida entre a relação de uma variável dependente categórica a um conjunto de variáveis independentes. Em suma, analisa diferentes aspectos ou variáveis de um objeto para depois determinar a classe na qual ele se encaixa melhor (GONÇALVES, 2013, p. 139-160). Essa técnica atribuiu um coeficiente para cada atributo, com base na influência desse sobre a variável alvo (ADEODATO, 2014, p. 891). Uma vez que o classificador a ser desenvolvido é multiclasse, utilizou-se a estratégia *one-vs-all* para a regressão logística.

Para a escolha dos três algoritmos acima, foram adotadas as seguintes premissas: não são permitidos valores nulos para os conteúdos dos arquivos; e os atributos da coluna *target* são categóricos, dentro de uma das sete classes constantes da Tabela 4.

3.4.2 Gerando um Design de Teste

Por se tratar de um problema de classificação, decidiu-se usar a acurácia entre os valores preditos e aqueles rotulados a fim de mensurar a qualidade dos algoritmos e operações para o problema em questão (GOLDSCHMIDT, 2005, p. 67). A acurácia foi calculada dividindo-se o número de registros classificados corretamente dividido pelo número de registros totais da base de validação.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Foi feita a divisão dos dados rotulados em dois conjuntos: treino (com setenta e cinco por cento dos documentos) e validação (contendo o restante), cujos percentuais foram definidos após uma série de tentativas, visando conciliar performance computacional com a acurácia alvo de 90%, conforme detalhado na seção 3.3.4 deste trabalho. Também foram ajustados hiperparâmetros nas funções e algoritmos usados, que serão detalhados na próxima seção.

3.4.3 Construção do Modelo

A fim de implementar o modelo classificatório para documentos referentes a procedimentos licitatórios, foram selecionadas as seguintes situações abaixo

Tabela 6 - Situações testadas para cada modelo

Situação	Descrição
1	500 Caracteres Relativos ao Conteúdo dos Arquivos
2	1000 Caracteres Relativos ao Conteúdo dos Arquivos
3	2000 Caracteres Relativos ao Conteúdo dos Arquivos
4	4000 Caracteres Relativos ao Conteúdo dos Arquivos

Fonte: Elaborada pelo autor (2020)

Para transformar os textos dos arquivos em formatos inteligíveis para os modelos citados na sessão 3.4.1, escolheu-se usar o método *CountVectorizer (CV)*, disponível no kit de ferramentas *scikit learn*.

A opção pelo CV deveu-se porque essa função converte o conjunto de textos analisados em uma matriz esparsa de contadores de palavras (CANDIDO, 2019, p. 1838), implementando uma representação *bag of words* de uma *string* ou de um arquivo (HACKELING, 2017, p. 51). Desse modo, era possível passar os conteúdos dos arquivos convertidos em matrizes numéricas, que são *inputs* válidos para os métodos *fit* e *transform* dos modelos selecionados. O método *fit,* grosso modo, calcula os parâmetros de aprendizado de cada modelo, salvando-os como um estado interno de objetos, enquanto o método *transform* aplica a transformação dos parâmetros definidos pelo *fit* ao conjunto de testes.

O *bag of words* é uma abordagem comumente usada em problemas de mineração de dados textuais, sendo uma forma de transformar dados não estruturados em um formato estruturado, especificamente em uma matriz na qual cada um dos documentos analisados será

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

representando como um vetor das palavras que estão presentes naquele documento (MATSUBARA, 2003, p. 8).

Também se utilizou o método do *TfidfVectorizer (Tfidf)* na construção do modelo. Esse método possui a característica de realizar uma transformação tf-idf (sigla em inglês para *term frequency-inverse document frequency*), podendo ser entendida como uma estatística numérica cuja finalidade é indicar o quão importante uma palavra ou um *token* é para um conjunto de textos (LESKOVEC, 2015).

Observou-se, contudo, que os resultados advindos do CV eram levemente superiores aos obtidos a partir do Tfidf, consoante ilustra a Tabela 7 abaixo, que usa os hiperparâmetros default ambos os modelos e sequências iniciais de quinhentos caracteres para cada arquivo analisado, de motivo pelo qual se decidiu pelo uso daquele método.

Tabela 7 - Comparativo entre CV e Tfidf

Modelo	Acurácia (%)	
	CV	Tfidf
Naive Bayes	87,65	80,26
Regressão Logísica	94,61	93,08
Floresta Aleatória	94,74	94,66

Fonte: Elaborada pelo autor (2020)

As simulações todas foram feitas considerando duas possibilidades distintas, a saber: (i) utilização de valores *default* para os parâmetros CV; e (ii) definição manual de valores para os hiperparâmetros "*max_features"* e "*ngram_range.*

O "*max_features*" (MF) define o tamanho máximo do vetor de características a ser utilizado pelo algoritmo (BONADIA, 2019, p. 54-55). Optou-se por três possibilidade: (i) parâmetros default (*"None"*); (ii) mil palavras – para os bigramas –; e (iii) duas mil palavras para os trigramas.

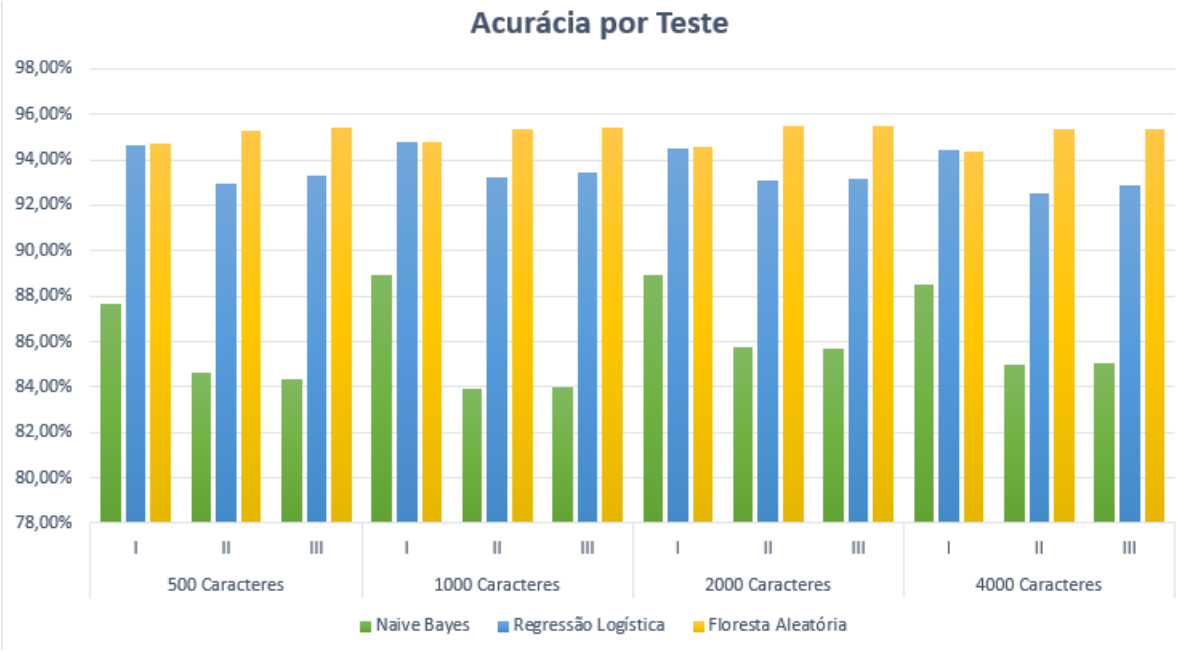
O "*ngram_range*" (NR) é o parâmetro que especifica o tamanho do conjunto de sintagmas - grupo de palavras que aparecem sequencialmente no texto (SANTOS, 2013, p. 21) - que serão usados pelo algoritmo. Eles são no formato de tupla (n,m), em que "n" representa o limite inferior e "m" representa o limite superior dos conjuntos de n-gramas extraídos pelo CV.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Usaram-se apenas unigramas (um único sintagma), bigramas (dois sintagmas) e trigramas (três sintagmas) no presente trabalho. Os unigramas (1,1) são os valores *default* no *scikit-learn*.

A Figura 12 ilustra a evolução dos resultados obtidos por algoritmo e hiperparâmetros. A situação I refere-se ao uso dos hiperparâmetros *default* do CV. A situação II reflete o conjunto MF 1000 e NR (1,2). A situação III, por sua vez, reporta os resultados de MF 2000 e NR (1,3).

Figura 12 - Resumo das acurácias obtidas



Fonte: Elaborada pelo autor (2020)

A Tabela 8 apresenta as combinações de hiperparâmetros que apresentaram a melhor acurácia em cada algoritmo. Além disso, o tempo de execução de cada execução foi mostrada para fins comparativos.

Tabela 8 – Melhores Resultados dos Modelos

Algoritmo	Quantidade de Caracteres	Hiperparâmetros		Acurácia (%)	Tempo de Execução (s)
		MF	NR		
Naive Bayes	500	None	(1,1)	87,65	0,1
	1000	None	(1,1)	88,95	0,2
	2000	None	(1,1)	88,94	0,4
	4000	None	(1,1)	88,47	0,5
Regressão	500	None	(1,1)	94,61	82

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Logística	1000	None	(1,1)	94,76	182
	2000	None	(1,1)	94,50	402
	4000	1000	(1,2)	94,47	439
Floresta Aleatória	500	2000	(1,3)	95,42	132
	1000	2000	(1,3)	95,44	204
	2000	1000	(1,2)	95,47	138
	4000	1000	(1,2)	95,34	252

Fonte: Elaborada pelo autor (2020)

Fica claro que o algoritmo que na maior parte das vezes trouxe valores com maior acurácia e menor variabilidade foi o da Floresta Aleatória, que consistentemente ultrapassou a acurácia de 95%.

O algoritmo de Regressão Logística também retornou bons resultados todavia as predições foram consistentemente piores nesse modelo ao se adotar os hiperparâmetros MF 1000 e NR (1,2). Nas demais situações, em especial com os hiperparâmetros default do CV, ficaram próximas àqueles retornados pelo da Floresta Aleatória. Destaca-se, contudo, que os tempos de execução das maiores acurácias nesse algoritmo foram consistentemente superiores aos da Floresta Aleatória, exceto no teste com quinhentos caracteres como conteúdo do arquivo.

O algoritmo *Naive Bayes*, por sua vez, em nenhum teste conseguiu atingir a acurácia desejada de 90%. Desse modo, não foi considerado como uma solução viável, a despeito de seu tempo de execução ser muito inferior ao dos demais.

A Tabela 8 e a Figura 12 demonstram que não necessariamente um maior conjunto de atributos refletirá em uma melhor acurácia para as predições das categorias dos documentos. Assim, concluiu-se que uma amostra dos quinhentos caracteres iniciais de cada arquivo é suficiente para atender à necessidade do acréscimo de um classificador de documentos no Alice. Nesse sentido, conclui-se que tanto o uso da Regressão Logística quanto da Floresta Aleatório são opções viáveis tecnicamente para a construção do classificador almejado, com vantagem para o primeiro modelo, em virtude de seu menor tempo de processamento.

A fim de entender melhor os resultados, fez-se uma matriz de confusão das predições constantes da situação III para o algoritmo Floresta Aleatória com 500 caracteres. Essa matriz está na Figura 13.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Figura 13 - Matriz de confusão

	EDITAL	TERMO DE REFERÊNCIA	MINUTA DE CONTRATO	MODELO DE PROPOSTA	DECLARAÇÃO DO LICITANTE	MINUTA DE ATA DE RP	ESPECIFICAÇÕES TÉCNICAS
EDITAL	3622	110	40	30	20	19	27
TERMO DE REFERÊNCIA	22	2766	1	5	4	1	61
MINUTA DE CONTRATO	12	25	1908	3	4	0	8
MODELO DE PROPOSTA	3	15	1	1435	3	0	21
DECLARAÇÃO DO LICITANTE	5	3	6	11	1414	0	9
MINUTA DE ATA DE RP	6	4	2	3	2	860	2
ESPECIFICAÇÕES TÉCNICAS	12	61	6	7	43	10	1047

Fonte: Elaborada pelo autor (2020)

3.4.3.1 Testes baseados em rotulação manual

A fim de se testar a acurácia do modelo em uma situação sem a prévia classificação automática de *targets* em função dos nomes dos arquivos, fez-se uma rotulação manual de cem arquivos texto. Para evitar um enviesamento da amostra, optou-se por diversificar o número de UASGs, bem como usar uma quantidade maior de arquivos do Poder Executivo (incluindo Administração Direta e Administração Indireta). Destaca-se, contudo, que os demais poderes também foram representados na amostra.

Esses arquivos foram rotulados manualmente de acordo com a Tabela 9.

Tabela 9 - Classificação manual da amostra de cem arquivos

Código Categoria	Nome Categoria	Quantidade
1	Edital	40
2	Termo de Referência ou Projeto Básico	21
3	Minuta de Contrato	7
4	Modelo de Proposta de Licitante	9
5	Declaração de Licitante	14
6	Minuta de Ata de Registro de Preços	7
7	Especificações Técnicas de Produtos	2

Fonte: Elaborada pelo autor (2020)

Esse conjunto de dados foi usado para validar a seguinte situação: Floresta Aleatória, 500 caracteres, MF 2 e NR (1,3) - o modelo de maior acurácia para 500 caracteres -, tendo sido obtida uma acurácia de 88%, cuja matriz de confusão é representada na Figura 14.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Figura 14- Matriz de confusão dos dados de validação

	EDITAL	TERMO DE REFERÊNCIA	MINUTA DE CONTRATO	MODELO DE PROPOSTA	DECLARAÇÃO DO LICITANTE	MINUTA DE ATA DE RP	ESPECIFICAÇÕES TÉCNICAS
EDITAL	40	0	0	0	0	0	0
TERMO DE REFERÊNCIA	0	21	0	0	0	0	0
MINUTA DE CONTRATO	0	0	7	0	0	0	0
MODELO DE PROPOSTA	0	0	0	6	0	0	3
DECLARAÇÃO DO LICITANTE	0	0	2	2	7	0	3
MINUTA DE ATA DE RP	0	0	0	0	0	7	0
ESPECIFICAÇÕES TÉCNICAS	0	2	0	0	0	0	0

Fonte: Elaborada pelo autor (2020)

Considera-se que o resultado foi satisfatório, pois apesar da amostra pequena a acurácia foi muito próxima daquela desejada. Além disso, algumas espécies de documentos foram classificadas sem nenhum erro, a exemplo de editais e termos de referências ou projetos básicos.

3.4.4 Avaliação do Modelo

Considerados tanto os testes com rotulação automatizada, quanto aquele feito com categorização manual, avalia-se que grande parte da os resultados obtidos no conjunto de treino e teste atendem à acurácia alvo de 90%, em especial o uso da Floresta Aleatória com o mínimo de quinhentos caracteres de conteúdo dos arquivos

Quanto aos dados do conjunto de teste com categorização manual, observa-se que o modelo mostrou ter alta confiabilidade para a classificação de editais e termos de referências (ou projetos básicos), que foram os elementos com maiores quantitativos (tanto na etapa de treino e teste, quanto na de validação), embora seja necessário refiná-lo para aumentar a confiabilidade de outras espécies documentais, menos comuns em procedimentos licitatórios, como por exemplo, declarações e modelos de propostas de licitantes.

3.5 Avaliação

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

3.5.1 Avaliação de Resultados

O modelo se mostrou útil para a classificação de editais e termos de referências/projetos básicos, que são os documentos mais comuns em certames licitatórios, responsáveis, entre outras coisas, por estabelecer as regras dos certames licitatórios, disciplinando sobre a participação no certame, sessão competitiva, recursos administrativos e justificativas para a contratação.

A utilidade para outras espécies documentais carece de evoluções e maiores quantidades de validação e treino, posto que foram identificadas particularidades que podem confundir o classificador. Um exemplo são certames que dividem os documentos em diversos arquivos, mas costumam juntar edital com o termo de referência, o que diminui o conjunto de documentos da categoria termos de referências a serem identificados, em especial porque a metodologia adotada foi buscar os primeiros 500, 1000, 2000 e 4000 caracteres, enquanto que na situação em tela os termos costumam vir como anexos – isto é, ao final – dos editais.

Feitas as considerações acima, propõe-se adotar o algoritmo de Regressão Logística com os hiperparâmetros *default* e quinhentos caracteres iniciais para a classificação documental, por ter sido um modelo de alta acurácia e baixo tempo de processamento durante os estágios de treino e validação da solução a ser implantada.

3.6 Aplicação

Não houve implantação do presente trabalho em ambiente de produção, tendo todos os testes e processamentos sido feitos em ambiente computacional não pertencente ao TCU.

Propõe-se, após a implementação das melhorias e evoluções sugeridas às seções 3.3.1 e 3.4.1 deste trabalho, implantar o projeto no ambiente de produção do TCU, além de integrá-lo ao Alice

4 CONCLUSÃO

Este trabalho apresentou o desenvolvimento de um modelo classificatório capaz de identificar a qual categoria de documento constante de um procedimento licitatório o arquivo pertence. Foram sete os tipos de documentos trabalhados: editais; termos de referências ou projetos básicos; minutas de termos de contratos; modelos de propostas dos licitantes;

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

declarações de licitantes; minutas de atas de registro de preço; e especificações técnicas de bens e serviços.

Para o treinamento do modelo, considerou-se um conjunto de 54.713 arquivos textos selecionados de licitações publicadas entre 2018 e 2019. Foram testados três algoritmos classificatórios: Floresta Aleatória, *Naive Bayes* e Regressão Logística, assim como algumas combinações de hiperparâmetros.

A melhor acurácia obtida foi com o uso do algoritmo Floresta Aleatória, com hiperparâmetros MF 2000 e NR (1,3), que apresentou acurácia de 95,47% na base de teste usando os primeiros dois mil caracteres dos arquivos a serem classificados. Todavia, considerando outros aspectos como o tempo de execução e o consumo de recursos computacionais, conclui-se que a adoção do algoritmo Regressão Logística com os hiperparâmetros *default* do CV e o conjunto inicial de quinhentos caracteres apresenta o melhor custo-benefício dentre todas as possibilidades analisadas.

Um teste adicional foi realizado considerando arquivos rotulados manualmente, obtendo uma acurácia de 88%, concluindo-se pela viabilidade da implementação do classificador de documentos licitatórios, em especial para a identificação e descobertas de certas classes de arquivos, como os editais de licitação.

Conclui-se, ainda, que um classificador de várias categorias documentais distintas necessita de uma ampla base de treino, a fim de robustecer os resultados desejados.

Quanto à realização de trabalhos futuros, sugere-se ampliar o rol de documentos a serem analisados (inserindo, por exemplo, planilha de custos estimados, cronogramas físicos e financeiros, memoriais de cálculo etc.), bem como refinar ainda mais os hiperparâmetros do modelo escolhido, otimizando a combinação desses.

Outra funcionalidade interessante, e capaz de agregar valor às análises dos AUFC do TCU, é de pegar um arquivo único e classificar cada parte de seu corpo textual em uma seção, dividindo blocos de *kilobytes* em análises específicas. Isso possivelmente permitiria a segmentação de documentos com vários anexos.

Por fim, uma evolução do modelo, e cuja necessidade originou o presente trabalho, é integrá-lo ao Alice, colaborando para a evolução e atualização do sistema.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

REFERÊNCIAS

ADEODATO, Paulo JL; SANTOS FILHO, Maílson M.; RODRIGUES, Rodrigo L. **Predição de desempenho de escolas privadas usando o ENEM como indicador de qualidade escolar**. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2014. p. 891.

ARAÚJO, T.; RODRIGUES, J. **Utilização de Dataflow para previsão de aceitação de respostas no fórum StackOverflow.com**. Revista de Engenharia e Pesquisa Aplicada, v. 3, n. 3, 30 ago. 2018.

AZEVEDO, Ana Isabel Rojão Lourenço; SANTOS, Manuel Filipe**. KDD, SEMMA and CRISP-DM: a parallel overview**. IADS-DM, 2008.

BARREIRA, Elisa da Conceição Marques. **População e Enriquecimento de Ontologias através de Web Scraping. 2014**. Tese de Doutorado.

BONADIA, Graziella Cardoso et al. **Contribuições para acelerar o aprendizado sobre a construção de uma máquina de classificação de sentimentos utilizando processamento de linguagem natural**. Dissertação (Mestrado no Programa de Pós-Graduação em Engenharia Elétrica e de Computação) –Universidade Estadual de Campinas, São Paulo, 2019.

BRAZ, Lucas M. et al. **Aplicando Mineração de Dados para Apoiar a Tomada de Decisão na Segurança Publica do Estado de Alagoas**. 2009.

CALIXTO, Kennet; SEGUNDO, Caetano; DE GUSMÃO, Renê Pereira**. Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar**. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2017. p. 1447.

CANDIDO, Antonio Leandro Martins; JÚNIOR, Corneli; FREITAS, Aislan. **Moderação inteligente de mensagens em ambientes virtuais de aprendizagem para alunos privados de liberdade**. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2019. p. 1838.

CHAPMAN P. et al. **CRISP-DM 1.0: Step-by-step data mining guide**. Disponível em: <https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72>. Acesso em: 1º de março de 2020

CHIARA, Ramon. **Aplicação de técnicas de data mining em logs de servidores web. 2003**. Tese de Doutorado. Universidade de São Paulo.

CRIMINISI, A. et al. Regression forests for eficient anatomy detection and localization in computed tomography scans. **Medical Image Analysis**, volume 17, issue 8, p.1293–1303, dez. 2013.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

DE ALVARENGA JÚNIOR, Wagner José. **Métodos de Otimização Hiperparamétrica: Um Estudo Comparativo Utilizando Árvores de Decisão e Florestas Aleatórias na Classificação Binária**. 2018. Dissertação (Mestrado no Programa de Pós-Graduação em Engenharia Elétrica) –Universidade Federal de Minas Gerais, Minas Gerais, 2018.

DE LUCCA, J. L.; NUNES, Maria das Graças Volpe. **Lematização versus Stemming**. USP, UFSCar, UNESP, São Carlos, São Paulo, 2002.

DOMINGOS, Pedro; PAZZANI, Michael. ***On the optimality of the simple Bayesian classifier under zero-one loss***. ***Machine learning***, v. 29, n. 2-3, p. 103-130, 1997.

DO NASCIMENTO, Rafaella Leandra Souza; DA CRUZ JUNIOR, Geraldo Gomes; DE ARAÚJO FAGUNDES, Roberta Andrade. **Mineração de Dados Educacionais: Um estudo sobre indicadores da educação em bases de dados do INEP**. RENOTE-Revista Novas Tecnologias na Educação, v. 16, n. 1, 2018.

GENUER, R. et al. **Random Forests for Big Data**. Big Data Research, Elsevier, p.28-46, 2017

GONÇALVES, Eric Bacconi; GOUVÊA, Maria Aparecida; MANTOVANI, Daielly Melina Nassif. **Análise de risco de crédito com o uso de regressão logística**. Revista Contemporânea de Contabilidade, v. 10, n. 20, p. 139-160, 2013.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data mining: um guia prático**. Gulf Professional Publishing, 2005.

HACKELING, Gavin. ***Mastering Machine Learning with scikit-learn***. Packt Publishing Ltd, 2017.

LAUREANO, Raul; CAETANO, Nuno; CORTEZ, Paulo**. Previsão de tempos de internamento num hospital português: aplicação da metodologia CRISP-DM**. RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação, n. 13, p. 83-98, 2014.

LEMOS, Eliane Prezepiorski; STEINER, Maria Teresinha Arns; NIEVOLA, Julio César. **Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining**. Revista de Administração-RAUSP, v. 40, n. 3, p. 225-234, 2005.

LOPES, Maria Célia Santos. **Mineração de dados textuais utilizando técnicas de clustering para o idioma português**. Rio de Janeiro: sn, 2004.

MATSUBARA, Edson Takashi; MARTINS, Claudia Aparecida; MONARD, Maria Carolina. Pretext: **Uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words**. Techinical Report, v. 209, n. 4, 2003.

MORO, Sergio; LAUREANO, Raul; CORTEZ, Paulo. Using data mining for bank direct marketing: An application of the crisp-dm methodology. In: **Proceedings of European Simulation and Modelling Conference-ESM'2011**. EUROSIS-ETI, 2011. p. 117-121.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.

NOGUEIRA, Diogo Rafael Pinto. **Agile Data Mining: Uma metodologia ágil para o desenvolvimento de projetos de data mining**. 2014.

OLIVEIRA, G.L. NETO, M.G.M. **ExperText: Uma Ferramenta de Combinação de Múltiplos Classificadores Naive Bayes**. In: Anales de la 4ª Jornadas Iberoamericanas de Ingeniería de Software e Ingeiería de Conocimiento, 2004, Madrid. v.1, p. 317-32

LESKOVEC, J., RAJAMARAN, A., ULLMAN, J.D., (2015), **Mining of Massive Datasets**. Cambridge University Press, Cambridge, United Kingdom, 2nd Edition, 2015.

SANTOS, Fernando Leandro dos. **Mineração de opinião em textos opinativos utilizando algoritmos de classificação**. Bacharelado (Bacharelado em Ciência da Computação) – Universidade de Brasília, Brasília, 2013.

WIRTH, Rüdiger; HIPP, Jochen. CRISP-DM: **Towards a standard process model for data mining. In*: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining***. *London, UK: Springer-Verlag*, 2000. p. 29-39.

Para verificar as assinaturas, acesse www.tcu.gov.br/autenticidade, informando o código 64350675.