Daniel Aguiar da Silva

Definição de um Classificador de Serviços Frequentes Contratados pela Administração Pública Federal por meio de Aprendizagem de Máquina

Brasília

2020

DANIEL AGUIAR DA SILVA

Definição de um Classificador de Serviços Frequentes Contratados pela Administração Pública Federal por meio de Aprendizagem de Máquina

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

REFERÊNCIA BIBLIOGRÁFICA

SILVA, Daniel Aguiar. **Definição de um Classificador de Serviços Frequentes Contratados pela Administração Pública Federal por meio de Aprendizagem de Máquina**. 2020. 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 DF. 200 fl.

CESSÃO DE DIREITOS

NOME DO AUTOR: Daniel Aguiar da Silva
TÍTULO: Definição de um Classificador de Serviços Frequentes Contratados pela Administração Pública Federal por Meio de Aprendizagem de Máquina
GRAU/ANO: Especialista/2020

_____

Daniel Aguiar da Silva
aguiars@tcu.gov.br

Ficha catalográfica

Silva, Daniel Aguiar da

Definição de um Classificador de Serviços Frequentes Contratados pela Administração Pública Federal por Meio de Aprendizagem de Máquina / Daniel Aguiar da Silva; orientador, Edans Flávius de Oliveira Sandes, 2020

50 p.

Monografia (especialização) – Escola Superior do Tribunal de Contas da União, Curso de Especialização em Análise de Dados para o Controle, Brasília, 2020.

Inclui referências.

1.Análise de Dados. 2. Mineração de Dados. 3. Aprendizagem de Máquina 4. Auditoria. I. de Oliveira Sandes, Edans Flávius. II. Escola Superior do Tribunal de Contas da União. III. Título.

DANIEL AGUIAR DA SILVA

Definição de um Classificador de Serviços Frequentes Contratados Pela Administração Pública Federal por Meio de Aprendizagem de Máquina

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.

Brasília, 26 de março de 2020.

Banca Examinadora:

_____

Prof. Edans Flávius de Oliveira Sandes, Dr.

Orientador

Instituto Serzedello Corrêa – TCU

_____

Prof.ª Saul Campos Berardo, Me.

Instituto Serzedello Corrêa – TCU

AGRADECIMENTOS

Primeiramente, agradeço a Deus pelo dom da vida e por todas as experiências de aprendizado e engrandecimento que tem me proporcionado.

Agradeço aos meus pais, Silfredo e Graça, por todo o amor e suporte que me tornaram o que sou e onde hoje cheguei. Sei que nada nesse caminho foi fácil e o exemplo de força e resiliência me são um verdadeiro guia e motivo de imensa admiração. Aos meus irmãos, Érica, Flávia e Rafael, pelo amor e carinho e a convivência, que sempre foi uma grande fonte de aprendizado.

A Maria Luiza, minha esposa, pela consciência e consonância nos objetivos de evolução pessoal e profissional, com o necessário suporte para o nosso desenvolvimento, incluindo a dedicação incansável à família.

Às minhas Maria Clara e Maria Fernanda, que do alto de sua inocência, não podem imaginar quão inesgotável fonte de amor e força são para este pai sempre apaixonado. E que força! Sua ternura e amor alimentam minha alma todos os dias para a eternidade.

Aos colegas da Selog, pela parceria e lições diárias de qualidade profissional e amizade, bem como apoio e incentivo prestados durante este curso.

Aos colegas da Pós-Graduação pelo companheirismo exemplar nas batalhas enfrentadas durante o curso.

Ao Edans, meu orientador, pela paciência e confiança depositados no decurso deste trabalho, frente às adversidades, que não foram poucas.

Ao ISC, pelo empenho em fornecer o curso e proporcionar um grande aprendizado.

RESUMO

Este trabalho apresenta modelos de classificação dos serviços frequentemente contratados pela Administração Pública Federal por meio de aprendizagem de máquina baseada em técnicas de mineração de dados. O trabalho, desenvolvido com base na metodologia CRISP-DM, aborda especificamente os serviços de Manutenção de Bens Imóveis; Manutenção de Máquinas e Equipamentos; Seguros em Geral; Segurança Ostensiva e Monitorada e de Limpeza e Conservação. A técnica de classificação supervisionada foi aplicada sobre as descrições de objetos licitatórios com base nas respectivas classificações da natureza da despesa, obtidas na base do Sistema Integrado de Administração de Serviços Gerais – SIASG. A adoção da abordagem de classificação binária permitiu a realização individualizada de preparação dos dados para cada tipo de serviço, especialmente em relação à etapa pré-processamento textual, assim como da utilização da técnica e Análise ROC para obtenção dos melhores resultados dos modelos produzidos. A avaliação final dos resultados levou à junção dos serviços de Manutenção de Bens Imóveis com os serviços de Máquinas e Equipamentos em único modelo, totalizando quatro os modelos produzidos. Sua integração com o Sistema de Análise de Editais e Licitações – Alice, automatizará parte do trabalho de triagem realizado por Auditores do Tribunal, proporcionando maior eficiência aos trabalhos de fiscalização realizados.

Palavras-chave: Análise de Dados. Mineração de Dados. Aprendizagem de Máquina. Auditoria.

ABSTRACT

We present classification models for five services frequently procured by the Brazilian Public Administration so that it can partially automate the screening work performed by auditors from the Tribunal de Contas da União – TCU, the Brazilian Audit Court. Based on CRISP-DM, we modeled four binary supervised classifiers for classifying the five services, once we merged two of them due to similarity. At the evaluation phase, applying ROC Analysis to assess the best performance, the models could achieve the business objectives.

Keywords: Data Analysis. Data Mining. Machine Learning. Auditing.

SUMÁRIO

1 INTRODUÇÃO

O desenvolvimento do poder computacional, tanto o de processamento como o de armazenagem de dados, e o surgimento de técnicas de mineração de dados, bem como de aprendizagem de máquina, tem possibilitado a exploração de volumes de dados cada vez maiores, assim como permitido ao computador realizar inferências, de modo a incrementar a automação.

Em contexto global de crescente competitividade e escassez de recursos, a necessidade de aumento da produtividade tem sido suprida pela utilização de técnicas de análise de dados nas mais diversas áreas do conhecimento.

Frente a esse panorama, é comum a realização de especulações sobre a substituição de humanos por máquinas, ante às expectativas de automação de tarefas (PATI 2017). O serviço público brasileiro, conquanto não se insira em um contexto de competitividade mercadológica, também recebe os impactos da escassez de recursos, bem como das cobranças por maior eficiência.

Nesse sentido, o Tribunal de Contas da União – TCU – tem investido recursos no desenvolvimento de ferramentas que tornem a sua atuação no exercício do Controle Externo mais eficiente. Entre as principais ferramentas existentes no TCU, situa-se o Sistema de Análise de Editais e Licitações, denominado de Alice, desenvolvido em parceria com a Controladoria Geral da União – CGU.

Em síntese, o Alice realiza análises de editais com a finalidade de avaliar riscos de potenciais falhas com base em requisitos pré-definidos denominados tipologias ou trilhas de auditoria. Com base nessas tipologias, o Alice procede à classificação dos editais e direciona a análise das atividades de Controle a serem realizadas pelos Auditores do Tribunal.

As tipologias podem ser definidas com base em requisitos legais, da jurisprudência do Tribunal, ou mesmo em razão de aspectos de risco, materialidade e relevância relacionados com o objeto.

Os resultados produzidos pelo Alice são utilizados pelas áreas de controle externo do Tribunal, especialmente a Secretaria de Controle Externo das Aquisições Logísticas – Selog, que tem como um dos principais objetos de trabalho a fiscalização de procedimentos licitatórios.

O tema deste trabalho de conclusão de curso é tangente ao realizado pelo Alice, uma vez que se insere na análise de editais de licitação, por meio da classificação de seus objetos.

O objeto da licitação é o núcleo da contratação, uma vez que nele são descritos os elementos essenciais do serviço ou produto a ser contratado, dos quais se extraem os aspectos intrinsicamente relacionados à área do serviço/produto, sua relevância para o órgão contratante e riscos correlatos, bem como identifica-se a sua pertinência em relação ao plano de ações de controle.

Atualmente, a análise dos objetos é majoritariamente realizada de forma manual pelos auditores que recebem os relatórios do Alice, oportunidade em que as análises de risco, materialidade e relevância são complementadas com o objetivo de selecionar os procedimentos licitatórios sobre os quais haverá ação de controle.

A automatização da identificação dos objetos, além de economizar tempo dos auditores, pode possibilitar a automatização de uma série de análises ulteriores, como a análise dos fatores de risco inerentes ao objeto em confronto com os aspectos específicos da legislação e da jurisprudência do TCU.

Adicionalmente, uma vez que o novo projeto de lei, o PL 1292/95 (BRASIL 1995), de licitações em contratos prevê o sigilo sobre a estimativa de preços feita pela Administração, o que inviabiliza a sua análise por meio da sistemática atual do Alice, qual seja, a análise editalícia, a avaliação do objeto licitatório pode suprir alternativa de avaliação de risco e relevância. O Projeto de Lei está em fase avançada no Congresso Nacional, em etapa final de apreciação, pela casa iniciadora, das modificações realizadas pela casa revisora. Uma vez concluída esta etapa, seguirá à sanção presidencial.

Frisa-se, contudo, que o sistema Alice já conta com modelo computacional de classificação de objetos para contratações de Tecnologia da Informação – TI, formulado pela Coordenação-Geral de Auditoria de Tecnologia da Informação – Cgati da Controladoria Geral da União – CGU. No âmbito do TCU, este modelo foi adaptado pela Secretaria de Gestão de Informações para o Controle Externo – SGI para consumo da Secretaria de Fiscalização de Tecnologia da Informação – Sefti.

Para o presente trabalho, o modelo em questão fora adaptado para aplicar-se às contratações de interesse da Secretaria de Controle Externo das Aquisições Logísticas – Selog, especialmente as relativas a manutenção de bens imóveis, manutenção de bens móveis e equipamentos, seguro, limpeza e conservação e segurança. Esses objetos foram escolhidos

devido à sua natureza comum e de necessidade continuada pelos órgãos da Administração Pública, com alta frequência, volume e materialidade, com cerca de 9,7% das contratações realizadas pela Administração Pública Federal, por meio de pregão eletrônico, em 2019.

A alta frequência de contratações implica em maior concentração de representações afetas a parcela desses serviços feitas junto ao TCU, que poderá usufruir de maior automatização em suas análises, além de possibilitar uma distribuição mais eficiente de temas por equipes, o que justifica a realização deste trabalho.

1.1 OBJETIVOS

O objetivo geral deste trabalho é a implementação de um classificador de objetos de serviços contratados pela Administração Pública Federal por meio de técnicas de mineração de dados textuais.

O classificador comporá ferramenta do Alice para fornecer informações em auxílio às análises da Selog na avaliação e seleção de objetos para fins de fiscalização.

O objetivo específico é a criação de um modelo classificatório para cada uma das seguintes naturezas de objetos: Manutenção e Conservação de Bens Imóveis; Manutenção de Bens Móveis e Equipamentos; Seguros em Geral; Segurança Ostensiva e Monitorada e; Limpeza e Conservação;

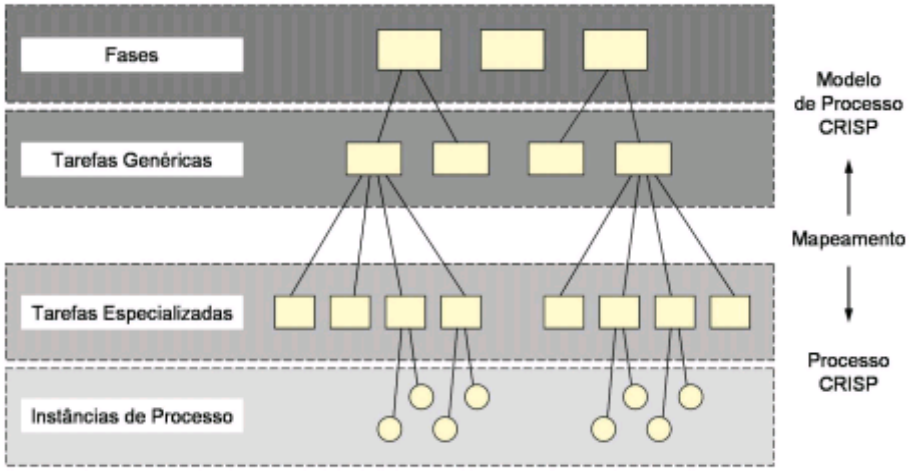
O restante deste trabalho está organizado da seguinte forma: a seção 2 abordará a metodologia CRISP-DM, na qual este trabalho se fundamenta; a seção 3 abordará o entendimento do negócio; a seção 4 abordará o entendimento dos dados disponíveis para a realização do trabalho; a seção 5 abordará a preparação dos dados; a seção 6 abordará a modelagem; a seção 7 abordará a avaliação dos resultados obtidos e; por fim, a seção 8 abordará as conclusões do trabalho.

2 METODOLOGIA

Este trabalho segue a metodologia definida por CHAPMAN et al (2000), denominada *Cross-Industry Standard Process For Data Mining* – CRISP-DM, que se baseia em abordagem hierárquica que consiste em um conjunto de tarefas descritas em quatro níveis de abstração a

saber: fase, tarefa genérica, tarefa especializada e instância do processo, conforme esquematizado na Figura 1, abaixo.

Figura 1– Estrutura hierárquica do CRISP-DM



Fonte: CHAPMAN et al (2000)

As fases organizam as tarefas genéricas, que têm como objetivo abranger todas as possibilidades de mineração de dados, provendo abrangência e estabilidade, de modo a cobrir todo o processo de mineração, independentemente do surgimento de novas técnicas. Já as tarefas especializadas são destinadas aos aspectos mais concretos das tarefas genéricas, com escopo mais específico.

A título de exemplo, em relação à fase de preparação dos dados, a tarefa genérica de limpeza dos dados pode ter tarefas específicas relacionadas à limpeza de dados numéricos e/ou limpeza de variáveis categóricas.

Por fim, a instância do processo representa a aplicação de todo o processo definido nos níveis mais abstratos ao caso concreto, com registro das ações, decisões e resultados alcançados no processo de mineração.

Embora a estruturação seja hierárquica, os autores afirmam haver apenas uma idealização de sequência de eventos.

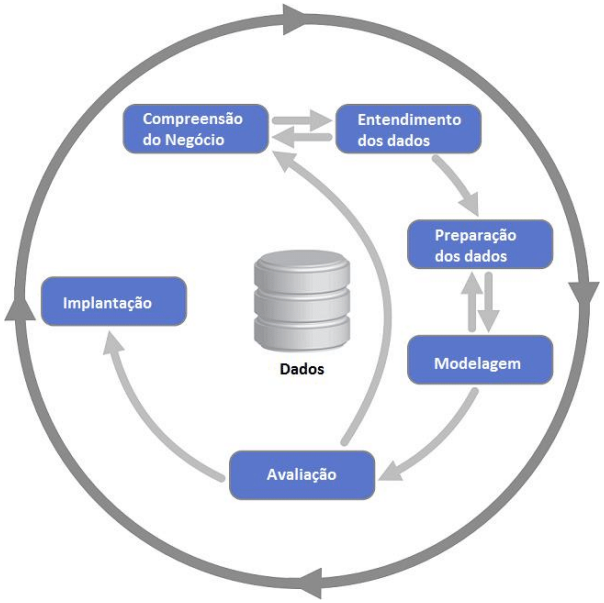
2.1 CICLO DE VIDA

O CRISP-DM prescreve um ciclo de vida composto por seis fases, de sequência flexível, e cujo fluxo é, necessariamente, de via dupla, com retroalimentação entre as fases,

definindo um processo iterativo. O produto de cada fase determina qual fase, ou tarefa em particular a ela pertencente, deverá ser realizada em seguida.

A Figura 2 ilustra as fases do processo, com as setas indicativas das principais e mais frequentes dependências entre elas. Por meio do círculo externo, a natureza cíclica da mineração de dados, que não finaliza quando da implantação da solução. Uma vez aprendidas novas lições, pode-se levantar questões de negócio mais específicas

Figura 2– Fases do Modelo de Referência do CRISP-DM



Fonte: Chapman et al (2000)

As fases acima esquematizadas são descritas da seguinte forma:

2.1.1 Compreensão do Negócio

Fase inicial para entendimento dos objetivos do projeto e dos requisitos a partir de uma perspectiva do negócio; definição do problema de mineração de dados; e elaboração de plano preliminar para alcance dos objetivos.

2.1.2 Entendimento dos Dados

Esta fase se inicia com a coleta dos dados e os respectivos procedimentos para aquisição de familiaridade com os mesmos, identificação dos problemas relativos à qualidade, bem como de aspectos que possam ajudar a levantar hipóteses quanto a informações subjacentes.

2.1.3 Preparação dos Dados

Tem como objetivo construir o conjunto de dados final para alimentar as ferramentas de modelagem a partir dos dados brutos inicialmente disponibilizados. As tarefas a ela associadas podem ser realizadas diversas vezes e sem ordenação específica.

2.1.4 Modelagem

Nesta fase ocorre as seleção e aplicação das técnicas de modelagem. Os parâmetros devem ser calibrados para os valores ótimos e a volta à fase de preparação de dados pode ser frequentemente necessária.

2.1.5 Avaliação

É a fase em que se busca assegurar que o modelo atende adequadamente aos objetivos do negócio. Tem como objetivo chave averiguar se objetivos importantes não foram considerados suficientemente, para concluir se os resultados de mineração poderão ser alcançados.

2.1.6 Implantação

É, de fato, a disponibilização do produto final para consumo do usuário. Este produto pode ser um simples relatório gerado ao final do processo, ou uma implementação complexa e repetível processo de mineração no empreendimento. O responsável pela implantação pode ser o próprio cliente.

3 ENTENDIMENTO DO NEGÓCIO

O uso do Sistema de Análise de Editais e Licitações – Alice – foi sistematizado pela Portaria TCU 296/2018. Em síntese, o Alice realiza o envio de e-mails aos Auditores do Tribunal de Contas da União com informações extraídas dos editais licitatórios a serem publicados, assim como de análises sobre os registros dos procedimentos licitatórios.

Em relação aos editais, o Alice realiza análises sobre os seus elementos em busca por aspectos que se enquadrem nas tipologias em que se baseia. Essas tipologias são, em essência, definições de requisitos que se deseja verificar, como uma determinada infração a algum dispositivo legal ou à Jurisprudência do Tribunal de Contas da União, ou elementos se enquadrem em critérios de risco, materialidade e relevância estabelecidos pelas Unidades Técnicas do Tribunal, por exemplo.

O trabalho aqui proposto se resume à análise da descrição dos objetos contidos nos editais de certames licitatórios, com a consequente classificação da contratação em um dos

seguintes tipos: 1) manutenção de bens imóveis; 2) manutenção de bens móveis e equipamentos; 3) seguros em geral; 4) limpeza e conservação e 5) segurança ostensiva e monitorada.

A fonte de dados é amparada nos artigos 14 e 40 da Lei 8.666/1993. O art. 14 da Lei reza que nenhuma compra será feita sem a adequada caracterização de seu objeto e indicação dos recursos orçamentários para seu pagamento, sob pena de nulidade do ato e responsabilidade de quem lhe tiver dado causa. Já o art. 40 da aludida Lei determina que a descrição do objeto seja precisa.

Considerando que a maior parcela dos serviços supracitados é considerada serviços comuns, pelo que devem ser realizados por meio de pregão eletrônico, é mandatório observar o mandamento do art. 3º da Lei 10.520/2002, que rege a referida modalidade licitatória, bem como o inciso IV do art. 30 do Decreto 5.450/2005, que a regulamenta.

Conforme o art. 3º da Lei 10.520/2002, a fase preparatória do pregão deverá observar, entre outras, a definição do objeto, que deverá ser precisa, suficiente e clara. Já o inciso IV do art. 30 do Decreto 5.450/2005 determina que deverá prever os recursos orçamentários com a indicação das respectivas rubricas. Conquanto o instrumento legal se utilize imprecisamente do termo "rubricas", a interpretação técnica conduz à classificação contábil da despesa orçamentária, especialmente referente à natureza da despesa.

A relevância do dispositivo retro dá-se pelo fato de a solução adotada neste trabalho se basear nas classificações orçamentárias de natureza da despesa, conforme será descrito na seção 3, referente ao entendimento do negócio. A associação da classificação orçamentária da natureza da despesa à descrição de objeto fornece o material suficiente e necessário à realização de análises com a finalidade da automatização classificatória.

3.1 Classificações Orçamentárias

O orçamento é o documento que reúne e dá publicidade a toda a programação do Poder Público por meio da estimativa de receitas e fixação das despesas que planeja realizar em um exercício. É, por excelência, o instrumento pelo qual o Parlamento controla o Poder Executivo, sendo, assim, instrumento de gestão, transparência e controle.

A estrutura do orçamento é definida com base nas regras da contabilidade pública, ramo da contabilidade aplicada aos órgãos da Administração Pública.

Devido à vastidão e complexidade do tema, o escopo desta seção se restringe aos conceitos estritamente necessários ao entendimento das classificações orçamentárias da natureza da despesa. Para maiores esclarecimentos, o Manual Técnico do Orçamento é disponibilizado, em versões anuais, pela Secretaria de Orçamento Federal do Ministério da Economia (BRASIL 2020).

A classificação da despesa é representada por um código alfanumérico definido com a finalidade de identificar informações qualitativas e quantitativas da programação orçamentária, que identificam, entre outros, o órgão responsável pelo gasto, os programas e ações das políticas públicas em execução e os montantes financeiros a eles alocados.

Em apertada síntese, a classificação da despesa tem como objetivo responder a questões como: quem é o responsável pelo gasto; em qual área a despesa será realizada; quais programas e objetivos serão implementados; o que e como será realizado; quais as metas e efeitos econômicos da sua realização; como serão aplicados os recursos; o que se pretende adquirir e; de onde virão os recursos.

Como o objetivo deste trabalho é a identificação do produto ou serviço previsto em edital, o aspecto da classificação da despesa relevante para o alcance dos objetivos é a natureza da despesa, mais especificamente, do elemento e subelemento nela identificados, que se referem ao que se pretende adquirir.

A classificação da natureza da despesa foi estabelecida pela Portaria Interministerial STN/SOF 163/2001 (BRASIL 2001) e é composta por um código de 8 algarismos, dispostos conforme a figura a seguir:

Figura 3- Codificação da Natureza da Despesa (MTO 2020)

1º	2º	3º	4º	5º	6º	7º	8º
Categoria Econômica	Grupo de Natureza da Despesa	Modalidade de Aplicação		Elemento de Despesa		Subelemento	

Fonte: MTO 2020 (BRASIL 2020)

O primeiro dígito, referente à categoria econômica da despesa, refere-se ao efeito econômico da despesa realizada, se influem ou não no patrimônio líquido da Administração. Possuem dois tipos: despesas correntes (código 3) e de capital (código 4).

As despesas correntes são afetas à manutenção dos equipamentos e serviços providos pela Administração e não contribuem, diretamente, para a formação de patrimônio.

As despesas de capital, por outro lado, contribuem, diretamente, para a formação de patrimônio da Administração, uma vez que se relacionam a modificações contábeis sobre bens e direitos, a exemplo da aquisição de máquinas.

O segundo dígito, referente ao grupo de natureza da despesa – GND – é uma subclassificação do objeto do gasto, composta pelos grupos relacionados na Tabela 1 abaixo:

Tabela 1 - Grupos de Natureza da Despesa

CÓDIGO	GRUPOS DE NATUREZA DA DESPESA
1	Pessoal e Encargos Sociais
2	Juros e Encargos da Dívida
3	Outras Despesas Correntes
4	Investimentos
5	Inversões Financeiras
6	Amortização da Dívida

Fonte: MTO 2020 (BRASIL 2020)

Os três primeiros grupos são afetos às despesas correntes, enquanto os três últimos são relacionados às despesas de capital. A codificação em comento tem como objetivo agrupar gastos em categorias consideradas relevantes para avaliação da composição alocativa dos recursos financeiros.

Os serviços selecionados para a realização deste trabalho são todos enquadrados no grupo 3 – Outras Despesas Correntes, uma vez que agrega todos os tipos de serviços contratados pela Administração.

O terceiro e quarto dígito são destinados a representar a modalidade de aplicação que, em apertada síntese indicam se a aplicação é realizada de forma direta pelo Ente detentor dos recursos, ou por meio de recursos recebidos por meio de transferência. Os objetos tratados neste trabalho são os realizados por meio direto de aplicação, código 90.

O quinto e o sexto dígitos são atrelados ao elemento do gasto, aspecto que inicia a identificação dos objetos. Os códigos de Elemento de Despesa foram instituídos pela Portaria STN/SOF 163/2001. Os tipos de contratações aqui abordados enquadram-se no Elemento de Despesa 39 – Outros Serviços de Terceiros – Pessoa Jurídica.

Deste modo, o padrão comum para os seis primeiros dígitos das classificações dos serviços abordados neste trabalho de conclusão de curso é o 3.3.90.39.

Por fim, o sétimo e o oitavo dígitos da classificação da natureza de despesa representam o subelemento, aspecto que, de fato, identificará os tipos de contratações enquadradas no escopo deste trabalho, que, no âmbito da União, são assim classificados:

Tabela 2 - Subelementos da Despesa tratados neste trabalho

Código	Descrição
33903916	Manutenção e Conservação de Bens imóveis (Mnt. Imóveis)
33903917	Manutenção e Conservação de Máquinas e Equipamentos (Mnt. Equipamentos)
33903969	Seguros em Geral (Seguros)
33903977	Vigilância Ostensiva / Monitorada (Vigilância)
33903978	Limpeza e Conservação (Limpeza)

A partir deste momento denominaremos o subelemento apenas como Objeto da Contratação, ou, simplesmente, Objeto.

3.2 Determinando os objetivos do negócio

O objetivo deste trabalho é classificar objetos licitatórios de serviços frequentemente contratados pela Administração Pública Federal para subsidiar as áreas usuárias do Alice, especialmente a Secretaria de Fiscalização das Aquisições Logísticas e seus Auditores, com informações automatizadas das classificações dos serviços contratados pela Administração Pública Federal.

3.3 Avaliação da Situação

O TCU possui a base do SIASG, que reúne os dados de editais licitatórios com informações desde o ano de 2012, especialmente os dados referentes aos pregões eletrônicos, principal modalidade licitatória para os serviços definidos no escopo deste trabalho, como mencionado anteriormente.

Adicionalmente, o Tribunal fornece infraestrutura e ferramentas para a realização do trabalho em ambiente denominado Labcontas, no qual o modelo resultante poderá ser implantado e executado diariamente.

3.4 Determinação de Objetivos de Mineração de Dados

Os modelos classificatórios obtidos deverão apresentar percentual de *recall* mínimo de 90% e precisão superior a 80%.

4 ENTENDIMENTO DOS DADOS

A fonte de dados para a realização deste trabalho é o BD_SIASG, base de dados que reúne as informações relacionadas às aquisições logísticas da Administração Pública Federal armazenada no Labcontas.

O SIASG – Sistema Integrado de Administração de Serviços Gerais – foi instituído pelo Decreto 1.094/1994 (BRASIL 1994) e tem por finalidade integrar os Órgãos da Administração Pública Federal direta, autárquica e fundacional e inclui informações sobre: divulgação e realização de licitações; emissão de notas de empenho; registro de contratos administrativos; catalogação de materiais e serviços; cadastro de fornecedores.

Os dados necessários ao treinamento do modelo relacionam-se às informações de licitações, por meio da descrição do objeto licitatório, e das informações necessárias ao empenho, por meio do código de natureza da despesa.

O empenho é a etapa da despesa em que a Administração se compromete com a realização do gasto, realizando a destinação do recurso para futuro pagamento.

Uma vez que a realização de uma despesa depende da devida previsão orçamentária, o empenho deve identificar as informações correspondentes no orçamento. Essa identificação é realizada com a indicação do código da classificação da despesa, do qual faz parte a classificação da natureza da despesa.

Conforme o modelo Siasg, o empenho é composto por itens de empenho. Esses itens correspondem aos itens de compra e, em relação às classificações orçamentárias, são rotulados com o respectivo código de subelemento. As informações da natureza da despesa relativas à categoria econômica, GND, modalidade de aplicação e elemento de despesa, por serem comuns a todos os itens, são registradas no empenho. Assim, a uma dada compra, corresponde um ou mais empenhos e os seus itens de empenho, conforme ilustrado na Figura 4:

Figura 4 - Relacionamento entre as entidades do modelo

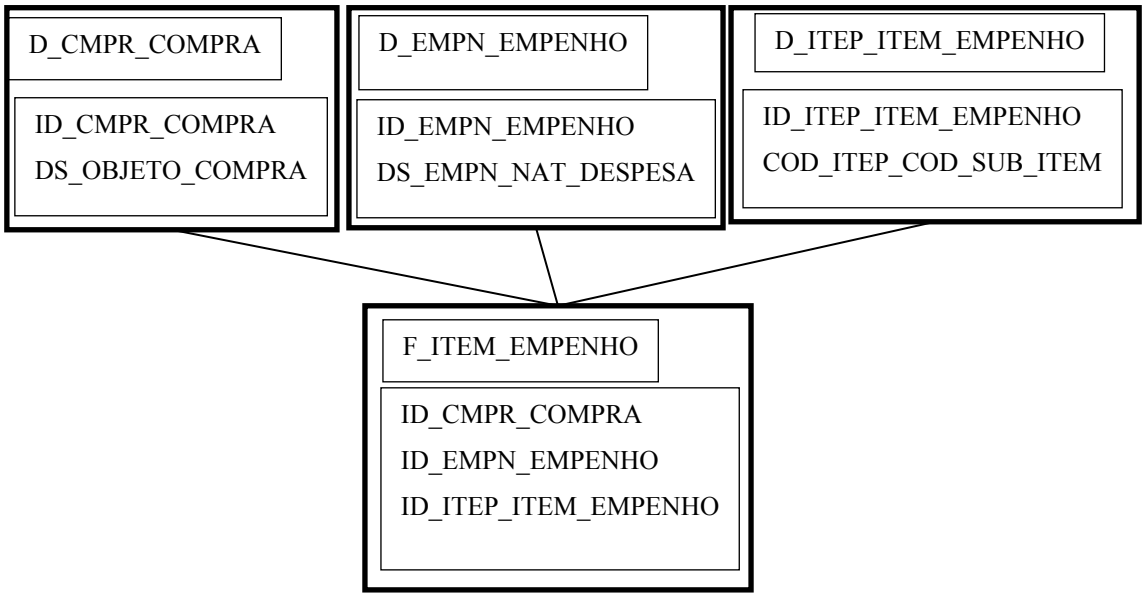
A implementação da relação representada acima é realizada por meio das tabelas D_CMPR_COMPRA, D_EMPN_EMPENHO, D_ITEP_ITEM_EMPENHO e F_ITEM_EMPENHO.

Em síntese, as tabelas que se iniciam com a letra ‘D’ representam ocorrências das entidades representadas no modelo da Figura 4, como a compra, o empenho e o item de empenho.

Todavia é a tabela F_ITEM_EMPENHO que contém as informações fáticas que correlacionam as compras aos empenhos e seus itens. Assim, ela informa quando houve ocorrência de empenho de um determinado item no âmbito de uma determinada compra. A junção dessas tabelas permite a recuperação dos detalhes de compra, empenho e itens de empenho.

A representação dessa estrutura é representada pela Figura 5:

Figura 5 - Relação das tabelas Siasg



4.1 Coletando os Dados

Uma vez que os dados utilizados para treinamento do modelo encontram-se na base supracitada, a coleta de dados será realizada na forma de consulta SQL.

A base de dados possui 1.498.865 registros de compras no período de 1/12/2012 a 20/2/2020. Considerando todas as ocorrências de itens empenhados na totalidade das compras ocorridas neste período, a base possui 19.181.265 registros.

Para a realização dos experimentos a coleta foi realizada de modo a compor dois segmentos em conformidade com as duas visões supracitadas, uma com base nas compras e outra com base nos itens de empenho.

A coleta realizada com base em itens de empenhos gera replicações das descrições dos objetos de compra, conforme ilustrado na Tabela 3. Essa replicação tem efeito de considerar a relevância, em termos de quantitativos de itens por contratação, reforçando as informações referentes ao tipo de objeto da contratação e fornecendo, maior insumo para a mineração.

Tabela 3 – Exemplo de replicação de descrição de objeto de compra com três itens de empenho

DESCRIÇÃO DO OBJETO
Este Pregão tem por objeto a contratação de empresa, visando o fornecimento de elevadores...
Este Pregão tem por objeto a contratação de empresa, visando o fornecimento de elevadores...
Este Pregão tem por objeto a contratação de empresa, visando o fornecimento de elevadores...

4.2 Descrevendo e Explorando os Dados

Conforme visto acima, o período de 1/12/2012 a 20/2/2020, a base do Siasg registra o total de 1.498.865 contratações. Dessas, 138.796 são afetas aos serviços objetos deste trabalho, cerca de 9,3% do total.

Inicialmente, são listadas 754 naturezas de despesa com o subelemento. Todavia, 25 são registradas com menos de 8 dígitos, pelo que se considera erro de registro na classificação da despesa realizada. Ao todo, essas 25 naturezas envolvem 639 contratações, volume inferior a 0,0005%. Uma inspeção em parte dessas contratações indica que o erro de cadastro teria sido nos subelementos cujos códigos são inferiores à dezena, pela ausência do dígito ‘0’ (zero) no cadastro dos códigos entre ‘01’ e ‘09’. O erro, entretanto, além de estatisticamente irrelevante nos dados, não ocorre nos serviços aqui selecionados, todos com código acima da dezena.

Eventual erro de enquadramento por parte dos registros não será considerado erro de digitação para exclusão de dados, mas terá tratamento no âmbito do “treinamento do modelo”, como será visto adiante.

Observou-se, ainda, haver 166 contratações realizadas na modalidade de aplicação 91, sendo 162 afetas à natureza da despesa ‘33913917’ – Manutenção de Máquinas e Equipamentos; 3 à natureza ‘33913978’ – Limpeza e Conservação e; 1 à ‘33913969’ – Seguros em geral. Ou seja, os mesmos de parcela dos objetos deste trabalho, todavia, realizados por meio de aplicações diretas decorrentes de operações entre Órgãos, Fundos e Entidades integrantes dos Orçamentos da Administração Pública Federal. Uma vez que somente muda o

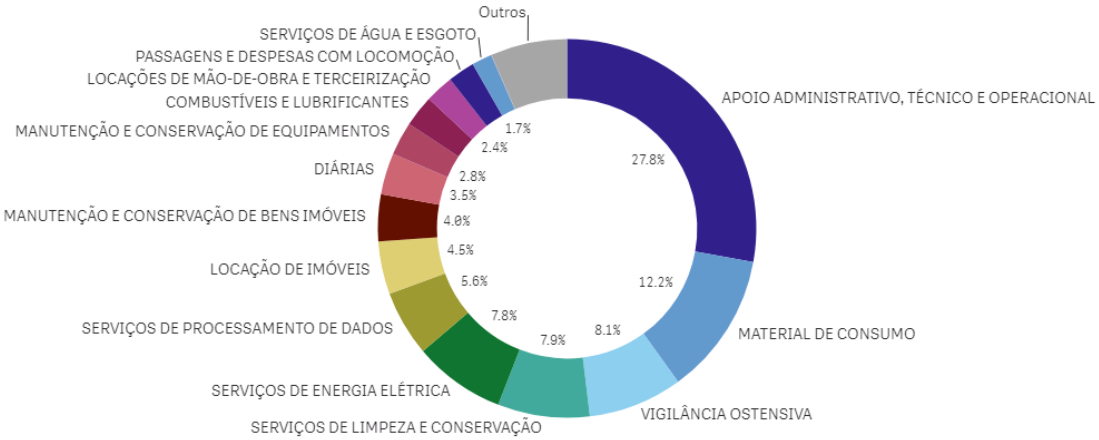
código da modalidade de aplicação, esses grupos deverão ser incluídos ao grupo originalmente definido, cuja Modalidade de Aplicação da classificação de natureza da despesa é 90.

Os dados levantados revelam que o grupo de manutenção de máquinas e equipamentos (‘33903917’) é o segundo em volume de contratações, com 67.894, ficando atrás somente das 88.691 contratações relacionadas a serviços de seleção e treinamento.

O grupo relacionado à manutenção de bens imóveis é tipo de contratação com o sexto maior volume, com 42.316 contratações. As contratações que o antecedem em volume são relacionadas a aquisição de materiais de consumo relacionadas a materiais para manutenção de bens imóveis e instalações (‘33903024’), ou seja, materiais que poderão ser utilizados em contratações de serviços de manutenção de bens imóveis; materiais de expediente (‘33903016’) e; materiais elétricos (‘33903026’).

A Figura 6 mostra a proporção de gastos por elemento de despesa obtido junto ao Portal de Custeio do Ministério da Economia, envolvendo diversos tipos de gastos correntes, e revela a importância material dos serviços selecionados neste trabalho, além da relevância da frequência de contratações, conforme apontado anteriormente.

Figura 6 - Proporção de gastos por elemento de despesa geral no exercício de 2019



Fonte: Painel de Custeio

A Figura 7 ilustra os percentuais dos subelementos envolvidos no escopo do trabalho e a Tabela 4 apresenta os valores correspondentes em reais.

Figura 7 - Proporção de gastos por subelemento de despesa no exercício de 2019

Fonte: Painel de Custeio e Siga Brasil

Conquanto os serviços de Seguros em Geral não representem proporção relevante dos gastos, sua inclusão no escopo deste trabalho teve motivação pelas necessidades levantadas no âmbito do Alice.

Tabela 4 - Valores de gastos por subelemento de despesa do exercício de 2019

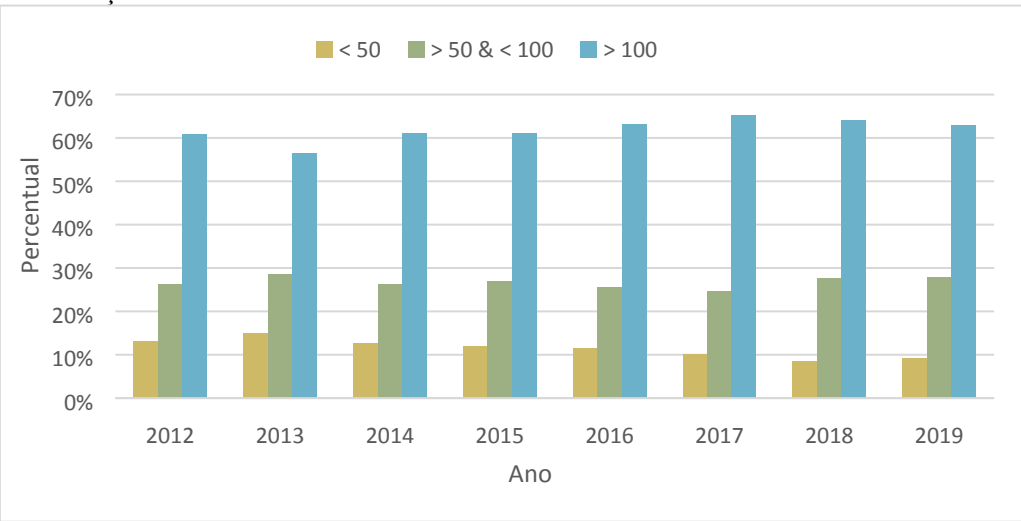
Subelemento de Despesa	Total Gasto no Item
VIGILÂNCIA OSTENSIVA	R$ 2.012.467.001
SERVIÇOS DE LIMPEZA E CONSERVAÇÃO	R$ 1.962.277.284
MANUTENÇÃO E CONSERVAÇÃO DE BENS IMÓVEIS	R$ 999.795.941
MANUTENÇÃO E CONSERVAÇÃO DE EQUIPAMENTOS	R$ 709.398.695
SEGUROS EM GERAL	R$ 273.129.008

Fonte: Painel de Custeio e Siga Brasil

Um dos aspectos que se buscou avaliar em busca de medida de qualidade para os dados da base do Siasg foi a completude das descrições dos objetos. Para tanto, levantou-se o quantitativo de caracteres das descrições dos objetos contratados no intuito de possibilitar a análise sobre sua influência nos resultados do modelo.

A Figura 8 apresenta gráfico com percentuais de comprimento das descrições dos objetos relativos à Manutenção de Bens Imóveis.

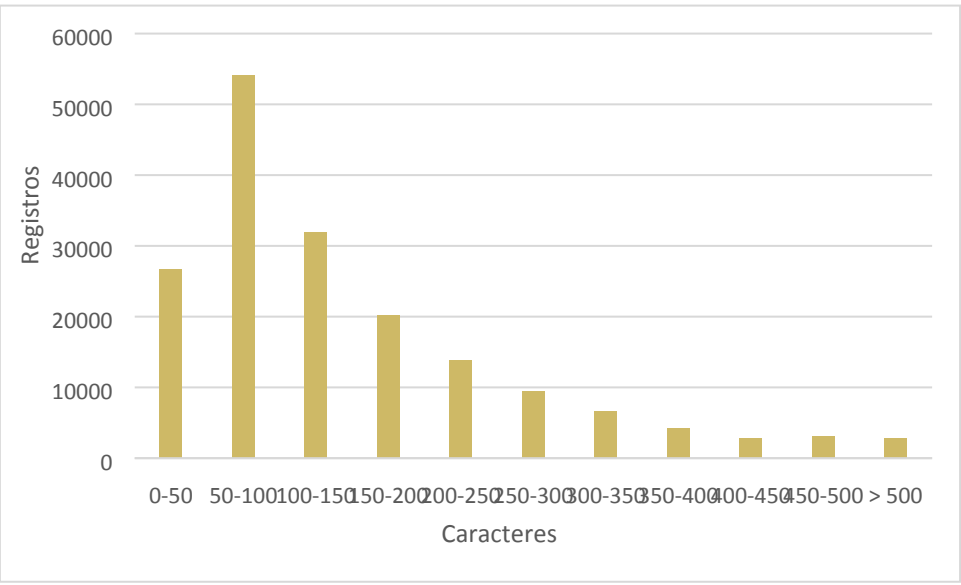
Figura 8 – Percentual de ocorrências por quantidade de caracteres usados na descrição dos objetos de Manutenção de Bens Imóveis



Nota-se que o número de caracteres utilizados nas descrições inicia uma leve tendência de aumento a partir de 2013. Descrições com mais de 100 caracteres passam dos 60% a partir de 2014. Já as descrições com menos de 50 caracteres ficam abaixo dos 10% a partir de 2017.

A **Error! Not a valid bookmark self-reference.** apresenta a distribuição de registros por quantidade de caracteres das descrições de objetos para o ano de 2019 de todos os cinco serviços e revela uma mais ocorrência de descrições com intervalo de 50 a 100 caracteres.

Figura 9 – Quantidade de registros por faixa de caracteres das descrições do objeto da Base SIASG do ano de 2019.



Considerando que quase todas as descrições se iniciam com o padrão "Objeto:", objetos cujas descrições possuem menos de 50 caracteres teriam, na prática, que descrever a contratação no máximo 43 caracteres. As duas hipóteses elencadas para a qualidade dessas

descrições são: a) descrições sucintas e objetivas, contendo elementos que identificam o subelemento, como “Manutenção predial”; b) descrições que não auxiliam na identificação do objeto da contratação e, assim, o subelemento a que pertence.

Conquanto inviável o levantamento estatístico para o enquadramento das descrições da base nessas hipóteses, ao consultar alguns exemplos, constatou-se que há, de fato, exemplos das duas hipóteses. Para a hipótese ‘b’, foram encontradas descrições como “Serviços”, “Outros serviços prestados”, além de abreviações, como “Srv. Mnt”, que não são identificáveis sem o auxílio de outras informações.

Todavia, o principal aspecto a ser observado é a fidedignidade das classificações dos objetos refletidas nos dados, conforme informações declaradas pelo gestor. Nesse contexto, foram avaliados manualmente 1.000 registros de classificações de subelemento de objetos no período de 2012 a 2014 e outras 1.000 para o período pós 2014. O erro encontrado para os objetos de manutenção de imóveis no primeiro período testado teve taxa de aproximadamente 15% na classificação. Os dados para o período posterior a 2014 apresentou taxa de erro caiu para cerca de 1,3% para a mesma quantidade de registros, indicando que os dados do segundo período estariam mais aptos a produzirem um modelo de maior qualidade.

Por fim, as Tabela 5 e 6 apresentam os quantitativos de registros de quantidade de compras e quantidade de itens por serviço. Nessas tabelas, a coluna “Percentual” indica a participação de cada serviço no total dos registros em questão. Esse total envolve todos os serviços e aquisições registrados na base. Ou seja, os demais somam mais de 90% dos registros, o que revela um conjunto de dados desbalanceado para as classes trabalhadas em relação ao conjunto toda da base selecionada. Note-se que na primeira o serviço referente a Manutenção e conservação de equipamentos (Mnt. Equip.), possui a maior relação percentual, com 4,5% das ocorrências. Já na segunda, o percentual cai abaixo da metade, ficando com 2,2%.

Tabela 5 – Registros de compras por serviço

Serviço	Quantidade	Percentual
Mnt. Imóveis	42.316	2,8%
Mnt. Equip.	67.894	4,5%
Seguros	11.447	0,76%
Vigilância	2.860	0,19%
Limpeza	14.279	0,95%

Tabela 6 – Registros de compras por itens de empenho

Serviço	Quantidade	Percentual
Mnt. Imóveis	787.502	4,1%
Mnt. Equip.	433.771	2,2%
Seguros	44.868	0,23%
Vigilância	25.211	0,13%
Limpeza	77.631	0,4%

5 PREPARAÇÃO DOS DADOS

Como sequência ao processo prescrito pelo CRISP-DM, apresentamos as etapas de preparação dos dados.

5.1 Seleção de Dados

Diante das análises realizadas na seção de exploração dos dados, a seleção de dados foi realizada em duas iterações, realizadas no intuito de refino dos dados para observar potenciais melhorias nos resultados dos modelos construídos. A primeira iteração considerou os dados a partir de 12/2012, enquanto a segunda utilizou dados a partir de 12/2014.

A seleção considerou as tabelas relativas a compras, empenho e item de empenho, considerando compras que possuam, ao menos, um item de empenho e originou dois conjuntos para a realização da modelagem, um baseado na seleção de objetos por item de empenho, e outro baseado apenas nas compras.

As tuplas selecionadas contêm a descrição do objeto e o código da natureza de despesa, incluindo o subelemento da despesa, que é o cerne da classificação empreendida neste trabalho.

Um aspecto particular da seleção dos dados refere-se ao conjunto de teste construído para teste do modelo. Ante os problemas relatados na comunidade de Inteligência Artificial em relação aos conjuntos desbalanceados (CASTRO 2011), que tendem a apresentar dificuldades em diferenciar grupos, favorecendo as classes de maior ocorrência, buscou-se produzir um conjunto de testes balanceado, com amostras aleatórias de igual tamanho para cada classe. Essa estratégia tem como objetivo possibilitar um melhor trabalho de refinamento do modelo.

Para cada classe foram selecionados mil objetos, totalizando, inicialmente, seis mil ocorrências, número que inclui mil elementos não pertencentes a quaisquer dos cinco serviços

objeto deste trabalho. Os dados em comento foram dissociados da base de treinamento e validação, de modo a serem utilizados especificamente na avaliação final dos modelos construídos.

Conduziu-se, na oportunidade, uma revisão manual das classificações realizadas originalmente pela própria Administração Pública, em nova rodada de exploração dos dados, na qual foi possível observar percentuais não desprezíveis de falhas nas classificações, conforme visto na seção de exploração dos dados. Entre falhas comuns observadas exemplificam-se as contratações de manutenção de elevadores, que ora os gestores classificam na natureza referente à Manutenção de Bens Imóveis (33903916), ora na de Manutenção de Máquinas e Equipamentos (33903917). Outro exemplo comum é o enquadramento de contratações de serviços para eventos nas contratações de seguros.

Diante das falhas detectadas, uma segunda iteração de seleção dos dados foi conduzida para avaliar diferenças na qualidade dos dados. Conquanto o índice de erros dos serviços de Manutenção de Bens Imóveis (33903916) tenha sido reduzido em seis vezes, o mesmo não pôde ser observado nos demais serviços, havendo casos de aumento do índice de erros. A Tabela 7Tabela 7 - Índices de erros de classificação orçamentária por tipo serviço em cada iteração a seguir apresenta os índices para os serviços em cada iteração de seleção.

Tabela 7 - Índices de erros de classificação orçamentária por tipo serviço em cada iteração

Serviço	1ª Iteração	2ª Iteração
Manutenção de Bens Imóveis (33903916)	18,4%	1,3%
Manutenção de Máquinas e Equipamentos (33903917)	4,5%	8,4%
Seguros (33903969)	1,1%	8,5%
Vigilância Ostensiva e Monitorada (33903977)	5%	4,9%
Limpeza e Conservação (33903978)	5,1%	3,1%

5.2 Limpando dos Dados

O padrão textual das descrições dos objetos da base do Siasg é conter, no início das descrições dos objetos, o termo “Objeto:”. Parte das descrições contêm, ainda, o termo “Pregão Eletrônico - ”, compondo a seguinte sequência: “Objeto: Pregão Eletrônico - ”.

O processo de limpeza adotado nesta etapa realizou a extração destes termos, quando presentes nos dados.

Em relação aos erros de classificação detectados quando da seleção da base de teste, optou-se por corrigi-los, tarefa realizada manualmente.

Conquanto haja expectativa de manutenção da taxa de erros na base de treinamento, a correção manual é impraticável, uma vez que possui quase 1,5 milhão de registros. A utilização de técnicas automatizadas, como a utilização de expressões regulares, para ser completa, envolve um esforço de revisão ampla dos registros.

Ademais, corre-se o risco de ineficiência e de ineficácia, uma vez que não há padrões claros nas construções textuais de descrições dos objetos licitatórios, assim como há grande número de aspectos dos serviços envolvidos nas contratações, o que poderia levar à necessidade de definição de grande número de expressões regulares para sua identificação.

Ainda, há casos de grande semelhança das descrições para tipos de contratações diferentes, como os de **serviços** de manutenção de bens imóveis (natureza 33903916) e a **aquisição de materiais** para manutenção de bens imóveis (natureza 33903024), ou mesmo entre os serviços de bens imóveis e os serviços de manutenção de máquinas e equipamentos, o que aumenta o nível de dificuldade e trabalho para a definição das expressões necessárias.

5.3 Excluindo Dados

Não houve exclusão de dados específicos, mas somente filtros relacionados às constatações obtidas na exploração da base, a exemplo da mudança do período de abrangência das contratações, que na primeira iteração se iniciou em 12/2012 e, na segunda iteração, se iniciou em 12/2014.

5.4 Construindo Novos Dados

Aos dados coletados, adicionou-se uma coluna, denominada “tipo”, referente à classificação do objeto.

Uma vez que os modelos classificatórios implementados neste trabalho são binários, ou seja, dado um serviço, o classificador deve identificar se é do tipo especificado ou não, a coluna “tipo” deve refletir a condição para o enquadramento ou não no serviço especificado, atribuindo o código ‘1’ caso a ocorrência pertença ao serviço e ‘0’, caso não pertença.

6 MODELAGEM

Em síntese, o problema deste trabalho é a extração de informações de dados não estruturados representados por textos produzidos no âmbito da Administração Pública Federal, com finalidade classificatória.

A técnica mais apropriada para o problema em questão é a mineração textual. Uma vez que a cada objeto corresponde uma classificação de natureza da despesa, a relação entre objeto e classe é unívoca, ou seja, a classificação deve ser simples.

Adicionalmente, a base de dados disponível possui as informações de classificação de modo a guiar o modelo, que deverá ser, portanto, supervisionado.

6.1 CLASSIFICAÇÃO TEXTUAL

A mineração textual é a aplicação de técnicas de mineração de dados a texto. A classificação textual, uma de suas aplicações, é a atividade de rotulagem de textos em linguagem natural com temas categóricos sobre um conjunto predefinido (SEBASTIANI 2002).

Uma etapa peculiar à mineração textual é a preparação dos dados para a execução do processo classificatório. Esta etapa, denominada pré-processamento, tem como objetivo preparar o texto para obtenção da melhor extração de informação.

Uma vez que o texto a ser classificado pode possuir caracteres e palavras que não apenas possuem baixo teor de qualidade informativa sobre seu conteúdo, a exemplo de caracteres especiais e preposições, que aumentam a complexidade do trabalho de classificação, deve passar por um tratamento prévio, denominado de pré-processamento, para redução dessa complexidade e melhor adequação à atividade classificatória.

A etapa de pré-processamento é composta por atividades como "tokenização", normalização e "stemização" (*stemming*). O resultado do processo é um conjunto de termos independentes, denominado *features*, normalmente organizado no formato de um vetor de palavras, ou *bag of words* – BOW.

A tokenização divide o texto em unidade, que podem ser números e termos, que podem ser palavras simples (unigramas), ou compostas (*e.g.* bigramas, trigramas, etc.).

A normalização envolve a conversão de letras e palavras maiúsculas em minúsculas, ou vice-versa, remoção de pontuação, caracteres especiais, espaços e as chamadas *stop words*, que são termos comuns da linguagem, cuja frequência de utilização não permite agregação de

valor de significado ao conteúdo textual. As preposições, a exemplo das palavras “de” e “para”, são exemplos normalmente enquadrados como *stop words* na língua portuguesa.

A *stemização* é a redução das palavras flexionadas ao seu radical.

A ferramenta utilizada neste trabalho para a realização dessas atividades de construção do vetor de palavras e pré-processamento é o *Scikit-Learn* (https://scikit-learn.org/).

6.1.1 Algoritmos Aplicados à Mineração Textual

Entre os algoritmos disponíveis para aplicação na solução a ser construída para a classificação textual supervisionada, encontram-se, por exemplo, SVM (*Support Vector Machine*), *Random Forest*, *Naive Bayes* e Regressão Logística.

O modelo SVM (CORTES 1995) baseia-se na utilização de um hiperplano para separação das classes. A definição do hiperplano para tal tarefa é definida com base nas margens, distâncias entre os pontos representantes das classes. A estratégia de escolha originalmente definida se baseia na margem máxima entre as classes, de modo a buscar minimização de possíveis falhas de classificação.

O *Random Forest* (BREINMAN 1999) é um algoritmo que se baseia na construção de árvores de decisão com as respectivas classificações definidas de forma não correlacionada umas com as outras, de modo que cada uma definirá um voto para a classificação do objeto. A classe definida ao final do processo é a que possui o maior número de votos, estratégia que busca mitigar falhas obtidas por decisões individuais.

O *Naive Bayes* (LEWIS 1998) é um algoritmo de classificação que se baseia na estimativa de probabilidades das classes. O cálculo da probabilidade de cada objeto definirá a classe a que pertence. O termo *Naive*, traduzido do inglês, ingênuo, dá-se pelo fato de desconsiderar qualquer correlação entre os atributos do objeto analisado.

A Regressão Logística é um modelo amplamente utilizado na estatística, especialmente conhecido pela aplicação na área de saúde, e recentemente extensamente estudado na área de aprendizagem de máquina devido à sua relação com o SVM (ZHANG 2003). Tem sido aplicada à classificação textual binária com desempenho considerado compatível ao do SVM.

6.2 MÉTRICAS DE AVALIAÇÃO DE CLASSIFICADORES

As métricas mais utilizadas na avaliação de classificadores são: acurácia; precisão; especificidade; e *recall*, revocação ou sensibilidade.

6.2.1 Acurácia

A acurácia é a fração das predições corretas sobre o conjunto de teste. Indica o desempenho geral do modelo, registrando o índice de acerto geral do classificador, o que inclui, indistintamente, os acertos em relação à classe positiva e à classe negativa.

É dada pela equação: (VP + VN) / (VP + VN + FP + FN).

Onde:

VP = Verdadeiros Positivos;

VN = Verdadeiros Negativos;

FP = Falsos Positivos;

FN = Falsos Negativos.

Os resultados alcançados podem ser adequados em classes que possuam níveis percentuais semelhantes de ocorrências. Contudo, casos em que há grande discrepância na ocorrência das duas classes pode não indicar uma avaliação correta do classificador.

Assim, seria possível que um classificador não detectasse qualquer ocorrência da classe positiva e ainda indicasse índice de acerto de quase 100%. Essa situação hipotética poderia ocorrer em casos nos quais a classe positiva apresentasse menos de 1% das ocorrências de um conjunto de teste. O modelo poderia identificar todas as ocorrências como negativas e apresentar acurácia superior a 99%.

6.2.2 Precisão

Mede a taxa de **acerto** das predições positivas realizadas pelo classificador. Ou seja, dentre as predições positivas feitas, quantas estavam corretas. É dada pela equação: VP / (VP + FP).

É mais relevante em relação às demais medidas em situações nas quais os falsos positivos são considerados mais prejudiciais do que os falsos negativos.

6.2.3 *Recall*, Revocação ou Sensibilidade

É a capacidade de o classificador tem de identificar como positivos os casos que realmente são positivos. Assim, do universo das ocorrências da classe positiva no conjunto testado, quantas o modelo teria identificado.

Objetivamente é a taxa de verdadeiros positivos e é calculada pela equação: VP / (VP + FN).

É especialmente relevante em situações em que os falsos negativos são mais nocivos ao sucesso do modelo do que a obtenção de falsos positivos. Uma alta taxa de recall faz aumentar a taxa de verdadeiros positivos, o que é desejável quando é mais relevante se detectar os casos positivos.

6.2.4 Especificidade

É, em síntese, o recall para a classe negativa. Assim, objetivamente, é a taxa de verdadeiros negativos, capacidade de o classificador identificar como negativos os casos realmente negativos e é dada pela equação: VN / (VN + FP).

6.3 DEFINIÇÃO DO MODELO

A definição do modelo a ser implementado levou em consideração uma análise comparativa inicial entre os algoritmos supracitados, de modo a avaliar a aptidão de cada um à tarefa de classificar a base de dados levantada junto ao Siasg.

Uma vez que a estratégia definida na modelagem é baseada em classificação binária, foram definidos cinco modelos classificatórios específicos para cada serviço, de modo a responder se uma descrição de objeto é pertencente à classe do serviço (*e.g.* classe Manutenção de Bens Imóveis), também denominada classe positiva, ou não.

O volume de dados inicial da base de treinamento e validação foi limitado a 90.000 descrições de objetos licitatórios para cada modelo, utilizando a seleção de dados por item de empenho, com proporção de 70% para os dados treinamento e 30% para dados de validação. A limitação deveu-se ao fato de que o algoritmo SVM não produziu resultados com volumes de dados superiores, com ocorrências de estouro de memória da máquina onde executava quando da manipulação de dados com mais de um milhão de registros.

O teste comparativo foi realizado para os serviços de Manutenção e Conservação de Bens Imóveis (33903916), devido à complexidade descritiva dada pela amplitude de serviços e variedade de vocábulos utilizados nas descrições dos objetos, observados na fase de análise de dados.

Conforme citado na seção de seleção dos dados, o balanceamento dessa base tem como finalidade a otimização do modelo. Sua utilização desde a fase preliminar de seleção possibilita maior rastreamento da evolução dos resultados obtidos com refinamento do modelo desde as fases iniciais.

A Tabela 8 mostra os resultados aproximados de acurácia, precisão e *recall* obtidos na execução para o cenário descrito acima. À exceção do Random Forest, que utilizou 200

estimadores, os demais algoritmos foram utilizados com valores padrão de hiperparametros. O vetor de palavras utilizado foi o CountVectorizer, com o hiperparâmetro min_df = 20.

Tabela 8 – Resultados obtidos por Algoritmo

Algoritmo	Acurácia	Precisão	*Recall*
SVM	57%	28,8%	50%
Random Forest	67,8%	78,6%	62,3%
Regressão Logística	74,5%	83,4%	70,1%
Naive Bayes	77,8%	80%	75,2%

Ante os resultados, a escolha do *Naive Bayes* como modelo foi chancelada para a solução adotada, especialmente pelo percentual de *recall* obtido.

Com o objetivo de melhorar os resultados buscou-se trabalhar sobre o vocabulário, etapa de pré-processamento à mineração, de modo a preparar a base para obtenção de melhores resultados. Conquanto os testes comparativos iniciais já tenham se beneficiado de algum refinamento de pré-processamento, faz-se necessário maior aprofundamento para obtenção dos resultados propostos.

A realização ampla e indistinta dos processos de normalização, remoção de *stop words* e *stemização* não levaram os modelos a melhores resultados. Contrariamente, em alguns casos os resultados obtidos foram piores.

Enfatizou-se, assim, a exploração dos hiperparâmetros das funções de construção do vetor de palavras (*bag of words* – BOW) disponibilizados pela biblioteca do *Scikit-Learn*, *CountVectorizer* e *TfidfVectorizer*.

As opções de pré-processamento que se mostraram mais efetivas foram as relacionadas à construção da BOW, especialmente aos hiperparâmetros “ngram”, “min_df” e “max_df”.

O parâmetro ngram define o que se chama de n-gramas, composição de termos que compõem o texto, onde n representa o número de termos. Assim, podemos ter unigramas (n=1), bigramas (n=2), trigramas (n=3) e assim por diante.

O parâmetro “min_df” tem como objetivo realizar a exclusão de termos de baixa ocorrência nos documentos, aqui representados pelas descrições de objetos. Assim, ele define a quantidade mínima de documentos em que os termos ocorrem para que possam ser considerados relevantes na análise (*minimum document frequency*).

O parâmetro “max_df” tem como objetivo realizar a exclusão de termos de alta ocorrência nos documentos (*stop words*), definindo um percentual máximo de ocorrência (*maximum document frequency*). Essa opção se mostrou particularmente interessante por se mostrar alternativa mais completa do que a seleção das melhores *features* (utilizando a função *selectkbest* da *Scikit Learn*), operação em que se delimita o número de termos usados na mineração aos de maior pontuação, ou seja, que mais teriam influência sobre os resultados.

Nos experimentos realizados, não foi possível delimitar um número de termos adequado a cada serviço. Todavia, com base nos respectivos percentuais de registros, como elencados nas Tabela 4Tabela 5 e Tabela 6, da seção de exploração de dados, foi possível calibrar testes com máxima frequência de termos em documentos, de modo a alcançar resultados condizentes com os percentuais de precisão e *recall* definidos nos objetivos de negócio.

Essa parametrização baseou-se na hipótese de que ao aumentar o rigor sobre o que seria alta frequência de ocorrência, amplia-se a lista de *stop words* para além dos termos triviais, como preposições e conectivos da linguagem. Ao limitar os percentuais de frequência a índices compatíveis com o percentual de participação da classe na base de treinamento, há uma tendência de que os termos da classe passem a ter uma maior influência no peso do conjunto de vocábulos disponíveis para o classificador, possibilitando um maior balanceamento de termos e, assim, melhores resultados classificatórios.

Como exemplo, os serviços de Seguros em Geral (natureza 33903969), que possuem percentuais de registros de 0,23% (relação por itens de empenho), a restrição de máxima frequência de documentos a 2% (max_df=0.02), como última etapa de refinamento, levou ao melhor resultado em relação aos índices de precisão e *recall*. Esse percentual mostrou-se adequado por forçar a redução significativa de termos ao ponto de aumentar a relevância dos registros relativos ao serviço em comento. Os índices de *recall* e precisão passaram de 96% e 89,3% para 97,2% e 90,9%, respectivamente.

Já em relação aos serviços de manutenção, tanto o Bens Imóveis como o de Máquinas e Equipamentos, a definição de máxima frequência não apresentou bons resultados. Uma hipótese para a diferença em relação aos demais modelos é a de que o volume de termos para esses serviços, que, por conterem uma vasta gama de objetos específicos, dificulta o nível de padronização alcançada pelos demais. Como exemplo, os serviços de seguros são aplicados, em sua maioria, a bens imóveis ou móveis, como automóveis, e a pessoas. Já os serviços de

manutenção são descritos de formas variadas e realizados em imóveis, partes dele, como em salas, portões, auditórios ou sistemas de eletricidade ou hidráulico.

Por fim, para os serviços de limpeza e conservação, os melhores resultados foram obtidos com os dados selecionados com base apenas nas compras, enquanto que com os demais, a base de treino que proporcionou melhores resultados aos modelos foi a baseada em itens de empenho.

A Tabela 9 apresenta as configurações de melhor resultado para cada modelo, com a respectiva base utilizada para treinamento e validação, se I – por Itens de empenho, ou C – por Compra. Utilizou-se a totalidade dos dados disponíveis em ambas, na proporção de 90% para treinamento e 10% para validação.

Tabela 9 - Relação de hiperparâmetros e base usados por serviço

Serviço	n_grams	min_df	max_df	Base (I/C)
Mnt. Imóveis	(1, 3)	20	*default* = 1.0	I
Mnt. Equip.	(1, 3)	20	*default* = 1.0	I
Seguros	(1, 3)	1000	0.02	I
Vigilância	(1, 3)	1000	0.2	I
Limpeza	(1, 3)	1000	0.02	C

O próximo capítulo aborda a avaliação dos modelos. Nele apresenta-se o conceito de curva roc, utilizada no refinamento final do modelo para a melhor configuração de resultados à luz dos objetivos de mineração e apresenta os resultados finais obtidos.

7 AVALIAÇÃO

Com vistas a avaliar os resultados obtidos pelos modelos construídos ante os critérios de sucesso do negócio e a eles melhor ajustá-los, utilizou-se da Análise ROC sobre as informações de *threshold* geradas pelos modelos implementados.

Threshold é um valor que pode ser determinado como uma pontuação ou probabilidade, a partir do qual a classe positiva é escolhida sobre a classe negativa.

Em geral, o valor associado ao *threshold* para classificações binárias é, por padrão, de 0,5 ou 50%. Ocorrências em que o valor de pontuação ou de probabilidade é superior a um desses valores são classificadas como positivas.

Todavia, esse valor pode não ser considerado ótimo devido aos objetivos de negócio, relacionando-se especialmente às relações de precisão e *recall* e, portanto, pode ser alterado para melhor calibragem do modelo.

Por exemplo, se, como no caso deste trabalho, o *recall* possui maior relevância em busca de diminuir a perda de verdadeiros positivos, o *threshold* pode ser calibrado com o objetivo de aumentar o percentual de recall.

Nesse diapasão, com base na análise ROC realizou-se o ajuste de *threshold*, em busca de um *threshold* ideal para as classificações dos serviços.

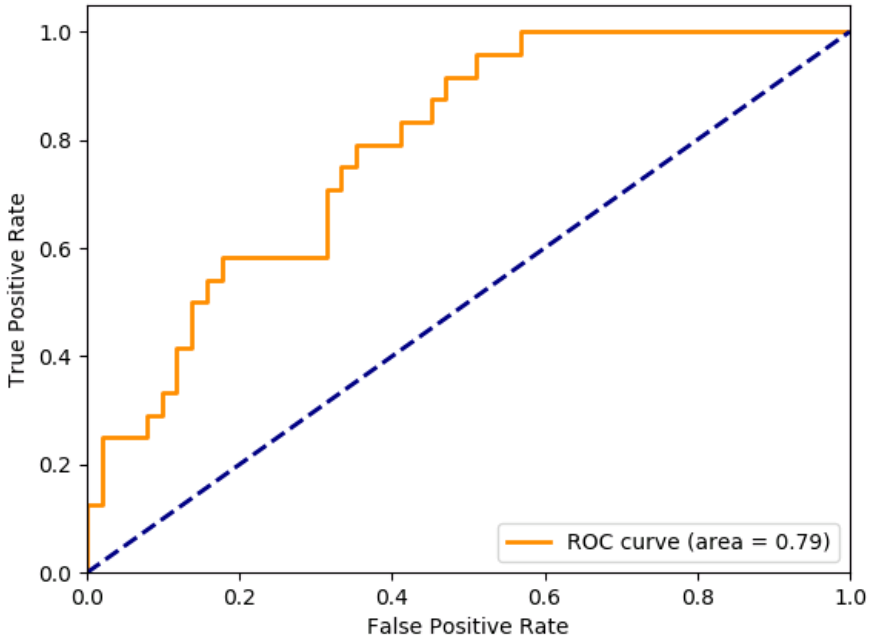
7.1 ANÁLISE ROC

A análise de característica de operação do recebedor, ou *Receiver Operating Characteristic* – ROC *analysis*, é uma estratégia popular para avaliação de desempenho de classificadores quando há duas classes (ZAKI 2014) e por meio da qual é possível realizar a identificação de um *threshold* ideal. Esta é a abordagem utilizada neste trabalho.

Esta técnica requer que o classificador defina pontuação (*score value*) para a classe positiva para cada ponto do conjunto de teste. A partir de então, com base nesse conjunto de pontuações, pode-se escolher o *threshold* de referência, ou ideal, para o modelo de classificação.

A escolha do *threshold* ideal é feita com base na análise gráfica sobre a chamada curva ROC, que reúne, para cada valor de *threshold*, as correspondentes taxas de falsos positivos, definidas sobre o eixo x, e as taxas de verdadeiros positivos, definidas sobre o eixo y.

Figura 10 - Exemplo de Curva ROC



Fonte: *Scikit-Learn*

A Figura 10 apresenta exemplo de curva ROC. Para cada par ordenado de Taxa de Falsos Positivos e Taxa de Verdadeiros Positivos, (x,y) respectivamente, plota-se o valor de *threshold* associado.

Considerando que o objetivo de todo modelo é obter a maior taxa possível de verdadeiros positivos e, corolário, a menor taxa possível de falsos positivos, o classificador ideal se posicionaria, portanto, no ponto mais alto à esquerda, (0,1), não apresentando falsos positivos, mas tão somente verdadeiros positivos.

Isso implicaria que um *threshold* de referência separaria todos os elementos da classe positiva dos elementos da classe negativa. Ou seja, o classificador atribuiria, a todos os elementos da classe positiva, pontuação superior aos da classe negativa. Por exemplo, todos os elementos da classe positiva receberiam pontuação superior a 0,6.

Ante a impossibilidade de se obter um classificador ideal, ou seja, em que o classificador pontua elementos da classe negativa acima dos da classe positiva, a maior proximidade deste ponto indica maior qualidade do classificador. Nesse sentido, em busca da otimização do modelo, deve-se buscar na curva o ponto que mais se adequar a esses requisitos

O exemplo da Figura 10 não apresenta um ponto claramente superior aos demais, de modo que a escolha deve envolver um *trade off* que terá influência sobre os resultados a serem alcançados pelo classificador.

Nesse diapasão, há alguns aspectos que se deve levar em conta, conforme os objetivos do negócio e de mineração, para a escolha do ponto mais adequado para um dado classificador, em busca de um resultado mais adequado ao negócio sobre o qual será aplicado. Na hipótese de ser mais relevante a identificação dos casos positivos, o *threshold* escolhido deve buscar pontos mais elevados da curva, admitindo, porém, um maior índice de falsos positivos.

7.2 AVALIAÇÃO DO THRESHOLD IDEAL PARA OS MODELOS DOS SERVIÇOS CONTRATADOS PELA ADMINISTRAÇÃO PÚBLICA

Para definição da estratégia de obtenção do *Threshold* ideal nos modelos construídos para classificação dos serviços aqui tratados, é necessário observar os aspectos mais relevantes ao trabalho desenvolvido no âmbito do Controle Externo em relação ao tema.

Atualmente a avaliação e classificação dos objetos são realizados por meio da leitura de um auditor, de modo intuitivo e informal, no volume total das ocorrências e tem como objetivo a avaliação de risco e relevância relacionados ao tipo de objeto envolvido.

A automação da classificação acarreta em ganho pelo só fato de, ao rotular um objeto, já oferecer um valor que só é inferido pelo auditor após a leitura, possibilitando a redução deste tempo, especialmente em casos nos quais já há um viés de busca por tipos de objeto. Ou seja, em caso de já haver um foco em certos tipos de objeto, o servidor já possui os objetos sem a necessidade de fazer toda a leitura e filtragem manualmente. De igual modo, a distribuição de processos no âmbito da Secretaria de Fiscalização interessada pode ser automatizada.

Nesse sentido, considerando o tempo gasto na leitura total para seleção e descarte dos objetos, a busca por um objeto erroneamente não incluído é potencialmente mais custosa do que o descarte de um objeto incluído indevidamente. Ou seja, modelo deve maximizar a detecção de ocorrências dos serviços objeto do trabalho, mesmo que isso signifique a inclusão de certo percentual de falsos positivos.

Assim, tem-se que a métrica mais relevante para este trabalho é o *Recall* ou Sensibilidade. Ou seja, a escolha do *Threshold* ideal na calibragem do modelo treinado com base nos dados do Siasg terá peso proporcionalmente considerado em caso de discrepâncias nos percentuais obtidos entre o recall e a precisão.

Uma vez que os percentuais de contratações de todos os serviços aqui tratados, em relação ao total de contratações a APF, é de 9,2% e, o serviço de maior volume de contratações, relativo a Manutenção de Máquinas e Equipamentos, é de somente cerca de 4,5%, a medida de acurácia poderá sofrer das distorções nos seus resultados, como visto anteriormente, de modo que esta medida de avaliação do modelo não será considerada para efeito de refinamento do modelo.

7.2.1 Algoritmo de avaliação do *Threshold* ideal

Em consonância com as considerações realizadas na seção acima, a avaliação de *thresold* ideal utilizada faz a busca pelo ponto da curva ROC que maximize a detecção dos serviços conforme as diretrizes inicialmente prescritas, quais sejam: Detectar 90% dos serviços (*recall*), com margem de ocorrência de cerca de 20% de Falsos Positivos nas predições.

Esta margem, embora seja um objetivo secundário, tem com intuito manter um conjunto conciso de predições para avaliação do auditor que consumirá as informações. Assim, o recall de 90% é objetivo primário e uma precisão de cerca de 80% é objetivo secundário.

Ante essa diretriz, o algoritmo implementado "percorre" a curva roc resultante da aplicação do modelo em busca do ponto cujo *Threshold* alcança os percentuais supracitados.

A Figura 11 abaixo ilustra o algoritmo implementado para a detecção do *Threshold* ideal. O método roc_curve invocado à linha 1, obtido no pacote do *Scikit Learn*, fornece as taxas de falso e verdadeiro positivos e de *thresholds* associados na forma de vetores.

Figura 11 - Algoritmo de detecção de *Threshold* Ideal

```
1    fpr, tpr, thresholds = roc_curve(...)
2    min_d = 1
3    min_i = 0
4    for i, t in enumerate(thresholds):
5        d = (1 - tpr[i]) ** 2 + fpr[i] ** 2
6        if min_d > d:
7            min_d = d
8            min_i = i
```

A variável min_d (linha 2) tem como objetivo fazer o controle do valor do cálculo realizado para definir o *threshold* ideal, juntamente com a variável min_i, que registrará o índice correlato no vetor de *thresholds* resultante do cálculo da curva roc.

O cálculo, definido na linha 5, é realizado sobre cada índice do vetor de *thresholds* (linha 4). Em síntese, o valor calculado é função da soma dos quadrados do complemento da

taxa de verdadeiros positivos, representada pela variável tpr, e da taxa de falsos positivos, representada pela variável fpr. Em outras palavras, é o quadrado da distância entre cada ponto da curva e o ponto (0,1).

Um aumento na variável tpr levará a uma diminuição do valor resultante da expressão $(1 – tpr)^2$. De igual modo um aumento na taxa de falsos positivos levará a um menor valor do resultado do quadrado da taxa de falsos positivos (fpr^2).

Se o valor encontrado for menor ao encontrado no índice anterior (linha 6), passa a ser o novo valor referência (linhas 7 e 8). O formato da equação permite o crescimento de ambas as taxas, no sentido de percorrer a curva do ponto inicial, inferior à esquerda, ponto (0.0, 0.0), ao superior à direita, ponto (1.0, 1.0).

Note-se que a primeira parte da equação, representada por $(1 – tpr)^2$, opera pela diminuição do valor resultante, enquanto o valor da parte representada por $(fpr)^2$ opera pelo aumento do valor à medida em que o algoritmo percorre os pontos da curva roc. Todavia, o valor só terá uma diminuição, passando a ser o ponto momentaneamente eleito, se houver um aumento proporcionalmente maior da variável tpr.

Caso haja algum ponto em que só haja aumento da taxa de falsos positivos, ou que essa taxa aumente em maior proporção que a de verdadeiros positivos, o resultado da equação será maior e, portanto, será descartado pelo algoritmo. Esse aspecto tem especial relevância quando a curva entra em sua parte superior, em que, sem evoluir no eixo y, passa a evoluir principalmente no eixo x, com valores de *threshold* que permitem ao classificador aumentar em demasia seu índice de erro.

A Tabela 10 a seguir ilustra os resultados obtidos pela aplicação exemplificativa do algoritmo. Os valores, que, por questão de espaço, contêm somente os quatro primeiros dígitos relevantes para fins meramente didáticos.

A iteração inicial é representada a partir do índice 122 do vetor (variável i), ocasião em que o valor da variável d será atribuído a min_d, uma vez que seu valor calculado, 0.01424, é inferior ao valor atual de min_d, 0.01470. O *threshold* momentaneamente eleito deixa de ser o do índice 117 e passa a ser o do 122.

Na iteração 123, o valor de tpr permanece inalterado, de modo que o algoritmo mantém o valor de min_d obtido na iteração anterior, 0,01424, uma vez que não há ganho de qualidade para o classificador. O aumento de fpr é descartado.

Na iteração 124 a variável tpr tem valor aumentado em relação à iteração anterior, o que, leva à mudança de min_d. O valor de fpr, inalterado, não influenciou os resultados.

Na iteração 125, somente o valor de fpr teve aumento. O valor de tpr permaneceu inalterado, de modo a manter o valor de min_d.

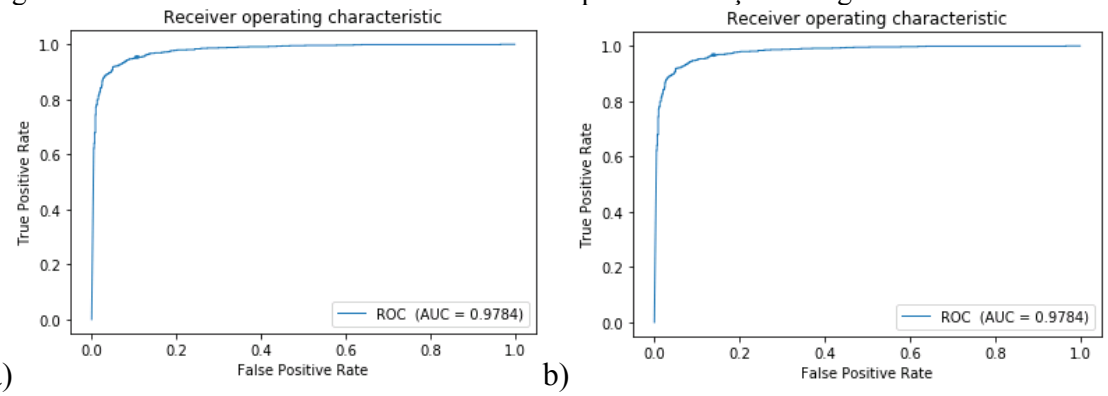
Por fim, nada obstante o valor de tpr tenha aumentado na iteração 126, o valor de fpr teve aumento em relação ao da iteração 124, fazendo com que o valor resultante de d seja superior ao calculado naquela oportunidade, mantendo inalterado o valor da variável min_d.

Tabela 10 – Valores de simulação do algoritmo

i	min_i	d	min_d	tpr	fpr
122	117	0.01424	0.01470	0.9101	0.07852
123	122	0.01439	0.01424	0.9101	0.07952
124	122	0.01422	0.01424	0.9111	0.07952
125	124	0.01454	0.01422	0.9111	0.08151
126	124	0.01437	0.01422	0.9120	0.08151

A Figura 12 ilustra, em gráfico, o efeito das mudanças realizadas pelo algoritmo. Após a mudança, o ponto escolhido está mais à direita da curva.

Figura 12 – Gráfico da curva roc resultante das etapas da execução do algoritmo



a) b)

É imperioso observar aqui o efeito de controle da equação definida na solução. O aumento da taxa de verdadeiros positivos, representado pela variável tpr, não pode vir acompanhado de índices maiores da taxa de falsos positivos (fpr).

A sensibilidade da equação pode ser calibrada, de modo a permitir que menores alterações nos valores de tpr e maiores em fpr possam resultar em alterações do *threshold*

escolhido, como ilustrado pela Figura 13, que apresenta alteração introduzida no algoritmo para multiplicar por 2 a expressão $(1 - tpr)^2$.

Figura 13 – Alteração do algoritmo para multiplicar o fator (1 - tpr)2 por 2

```
    fpr, tpr, thresholds = roc_curve(...)
    min_d = 1
    min_i = 0
    for i, t in enumerate(thresholds):
        d = 2*(1 - tpr[i]) ** 2 + fpr[i] ** 2
        if min_d > d:
            min_d = d
            min_i = i
```

Na prática, essa alteração permitiria um maior aumento da taxa de falsos positivos, para permitir avançar mais na taxa de verdadeiros positivos, o que pode ser determinante para permitir que o classificador alcance os percentuais definidos nos objetivos do trabalho.

A Tabela 11 apresenta, baseada no mesmo exemplo da Tabela 10, a simulação da execução do algoritmo alterado conforme a Figura 13 paras as iterações 124 a 126 com intuito de comparação. Os valores de d e min_d são, naturalmente, afetados com o novo fator de multiplicação, fazendo com que na iteração 126 as alterações dos valores de tpr e fpr que anteriormente não teriam afetado o valor de min_d, passem a afetar, alterando o valor do *threshold* escolhido.

As taxas de verdadeiros positivos e de falsos positivos que anteriormente estavam em 0.9111 e 0.07952, passam a ser 0.9120 e 0.08151.

Tabela 11 – Valores de simulação do algoritmo resultantes da alteração da Figura 13

i	min_i	d	min_d	tpr	fpr
124	122	0.02212	0.02231	0.9111	0.07952
125	124	0.02244	0.02212	0.9111	0.08151
126	124	0.02210	0.02212	0.9120	0.08151

7.2.2 Resultados obtidos

Inicialmente, deve-se registrar que foram realizadas diversas simulações para cada modelo envolvendo alterações dos hiperparâmetros de vetorização citados no capítulo 6, referente à modelagem, assim como a escolha da formação da base de treinamento utilizada, se baseada em itens de empenho, ou somente nas compras.

Ante os diversos resultados obtidos a escolha final dos modelos foi definida com base no resultado obtido após a análise roc, o que implicou na desconsideração de modelos com resultados preliminares melhores.

A tabela apresentada a seguir contém comparativo dos resultados sobre a precisão e *recall* dos modelos ante a identificação do *threshold* ideal, com respectivo fator de multiplicação do componente $(1 - tpr)^{2,}$ com base no algoritmo de avaliação da curva ROC apresentado acima.

Tabela 12 – Comparativo dos resultados alcançados com a identificação do *Threshold* ideal

Serviço	Acurária (%)		Precisão (%)		*Recall* (%)		Fator
Mnt. Imóveis	91,2	**93,0**	97,3	**91,8**	84,9	**94,5**	3
Mnt. Equip.	94,0	**94,2**	95,0	**92,7**	92,3	**95,3**	3
Seguros	94,7	**94,8**	90,9	**91,1**	97,2	**97,2**	3
Vigilância	84,9	**89,8**	92,7	**88,1**	74,3	**95,6**	3
Limpeza	92,2	**91,8**	89,6	**85,7**	93,6	**98,3**	3

Coincidentemente todos os serviços tiveram como fator mais adequado o 3. Outros valores também se mostraram adequados, a exemplo do fator 7 para o serviço de Manutenção de Bens Imóveis, cujo resultado de *recall* alcançou cerca de 95,4%. Inobstante, uma vez que a quantidade de falsos positivos superou a taxa de 10%, o fator 3 foi considerado o mais adequado para os objetivos de negócio.

Nada obstante cada modelo tenha alcançado os resultados mais abrangentes do negócio, um aspecto deve ser particularmente observado em relação aos dois modelos de manutenção, de Imóveis e Máquinas e Equipamentos.

O volume de termos coincidentes envolve termos cruciais para a caracterização das classes, como o próprio termo ”manutenção”. Por vezes, o termo ainda se faz acompanhar do objeto estrito da manutenção, a exemplo de “reator”, ou “auditório”, o que, ante uma falta de padrão ou frequência do termo, propicia a confusão classificatória.

Há ainda um volume de erros classificatórios de origem, quando o próprio gestor público confunde a classificação dos elementos de despesa. Nesse sentido, a análise dos dados revelou uma clara confusão do gestor em relação à classificação de manutenção de elevadores, que deveria ser relacionado com Manutenção e Conservação de Bens Imóveis (33339016). A

maioria das ocorrências encontra-se corretamente na natureza 33339016, todavia, há um número considerável de classificações realizadas na natureza 33903917 (Manutenção e Conservação de Bens Imóveis), o que faz aumentar o índice de falhas dos modelos.

Em testes específicos realizados com massa de dados envolvendo apenas as duas classes, considerando a de Manutenção de Imóveis como positiva, o índice de falsos positivos chegou a 68%. Gravemente, o número de falsos positivos da classe 16 superou o número de verdadeiros negativos.

Ao testar um modelo experimental baseado no treinamento sobre uma base composta apenas por objetos dos dois serviços em comento, não foram encontrados resultados significativamente melhores sobre a base de teste composta apenas por objetos desses serviços, com *recall* de 87,8% e precisão de 85,6%. Por outro lado, o modelo perde desempenho de modo significativo ao ser testado contra a base geral, que envolve outros tipos de contratações, ficando com *recall* de 72% e precisão de 71,2%.

Nesse diapasão, a solução de contorno adotada para o problema em questão foi a junção dessas duas naturezas em uma classe agregadora denominada serviços de manutenção de imóveis e equipamentos. Considerando o *threshold* ideal encontrado por meio da análise roc, o *recall* e a precisão para a classe positiva foram de 96,7% e 93,6%, respectivamente.

Há, ainda, contratações conexas, referentes às aquisições de materiais para essas contratações, a exemplo da 33903024 – Aquisição de Material para Bens Imóveis. Embora a observação sobre as ocorrências de erros do modelo revele registros de falhas de confusão classificatória por referir-se à aquisição de materiais para manutenção de bens imóveis, testes sobre o modelo foram realizados e concluiu-se que o índice de falhas na distinção entre as duas classes não é relevante, uma vez que a ocorrência de termos em comum, a exemplo do próprio termo "manutenção bens imóveis", é relativamente baixa.

Ante as considerações acima, o conjunto de modelos resultantes passa a ser de quatro e não mais de cinco modelos, uma vez que a junção dos serviços de Manutenção de Bens Imóveis e de Manutenção de Máquinas e Equipamentos, naturezas 33903916 e 33903917, respectivamente, implicou na eliminação dos modelos individualizados. A Tabela 13 abaixo apresenta os resultados finais de cada modelo.

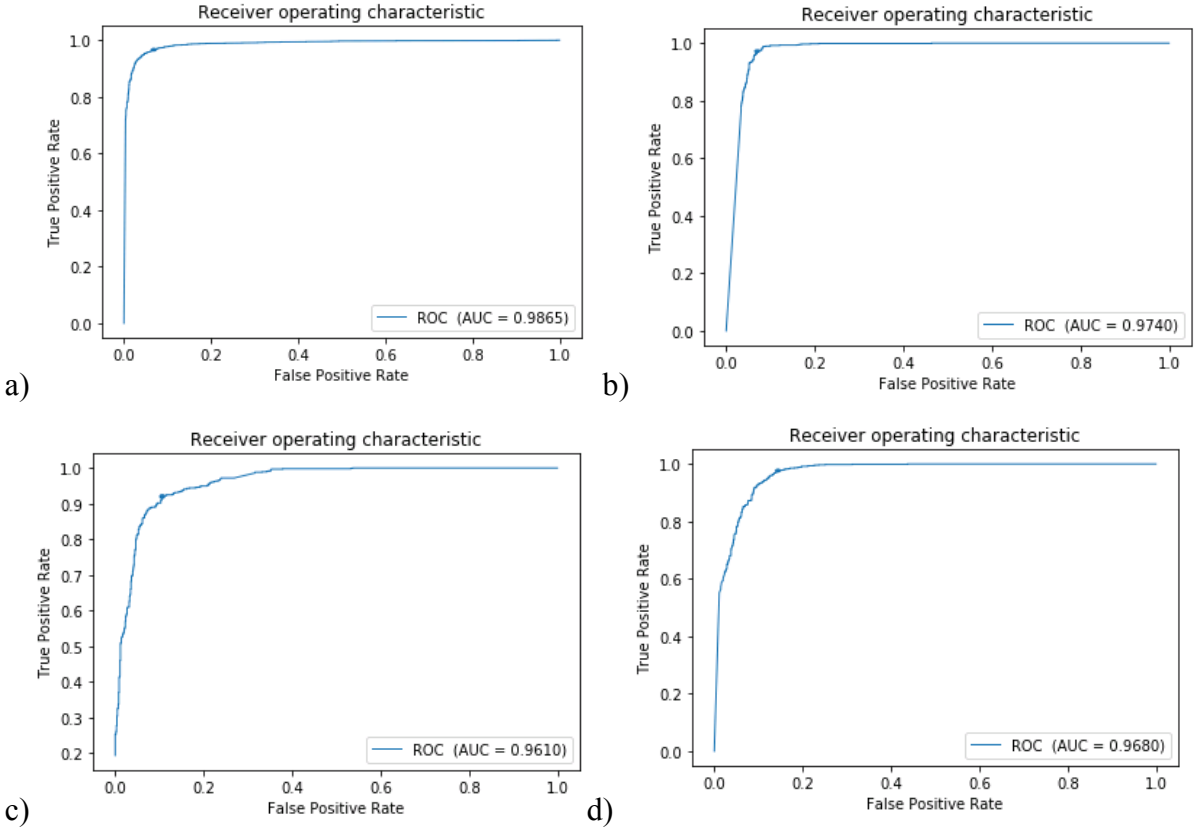
Tabela 13 – relação de modelos e respectivas métricas de avaliação

Serviço	Acurária (%)	Precisão (%)	*Recall* (%)
Manutenção	94,9	93,6	96,7

Seguros	94,8	91,1	97,2
Vigilância	89,8	88,1	95,6
Limpeza	91,8	85,7	98,3

A figura abaixo ilustra os gráficos de curva roc finais obtidos por cada um dos modelos. As curvas que seguem, da esquerda para a direita e de cima para baixo são: a) Manutenção; b) Seguros em Geral; c) Vigilância Ostensiva / Monitorada; e d) Limpeza e Conservação.

Figura 14 – Curvas roc resultantes para cada modelo



8 CONCLUSÃO

Este trabalho teve como objetivo a definição de modelo classificatório de serviços que são frequentemente adquiridos pela Administração Pública Federal por meio de aprendizagem de máquina.

Por meio de dados obtidos da base do SIASG – Sistema Integrado de Serviços Gerais, cinco modelos supervisionados, com base na classificação da natureza da despesa, foram treinados para identificar, nas descrições de objetos licitatórios, as classes referentes aos serviços de Manutenção e Conservação de Bens Imóveis (33903916); Manutenção e Conservação de Máquinas e Equipamentos (33903917); Seguros em Geral (33903969); Vigilância Ostensiva e Monitorada (33903977) e; Limpeza e Conservação (33903978).

Conquanto o objetivo inicial tenha se baseado na definição de um modelo classificatório por serviço, os resultados dos experimentos revelaram mais adequada a junção de dois dos cinco serviços, os de manutenção de imóveis e os de manutenção de equipamentos, em um único modelo. Assim, são quatro os modelos de classificação resultantes. Após ajustes do *threshold* por meio da curva ROC, os modelos obtiveram um recall variando de 95,6% a 98,3% e precisão variando de 85,7% a 93,6%.

Os resultados dos testes revelaram que os modelos construídos atendem aos critérios do negócio, de modo que, integrados ao Alice poderão prover informações mais completas e automatizadas às áreas de fiscalização do Tribunal de Contas da União, em especial a Secretaria de Fiscalização das Aquisições Logísticas – Selog.

Como trabalhos futuros, os modelos poderão ser integrados no sistema Alice e, à medida que os resultados de produção forem sendo coletados e avaliados, técnicas automatizadas de ajustes de hiperparâmetros poderão ser aplicadas no intuito de melhor refinamento dos modelos.

Adicionalmente, outros serviços e aquisições poderão ser igualmente modelados de modo a ampliar a classificação das aquisições logísticas. Nesse sentido, o processo aqui descrito deve ser replicado para outras naturezas de despesa. Outrossim, técnicas baseadas em redes neurais também poderão ser utilizadas como alternativa de evolução dos modelos.

Além disso, os modelos implementados possibilitam a implementação futura de funcionalidades que realizem análises de risco mais específicas aos serviços contratados. Para sua implementação, faz-se necessária a condução do levantamento dos principais riscos associados junto à área especialista, mapear nos editais as seções correlacionando-as aos riscos levantados e definir os alertas pertinentes. Um classificador textual poderá identificar os casos enquadrados nos padrões de riscos elencados.

Por fim, pode-se utilizar informações dos Acórdãos prolatados pelo TCU em relação aos serviços classificados para a avaliação de itens dos editais, de modo a identificar potenciais irregularidades e impropriedades conforme a Jurisprudência da Corte de Contas.

A partir de um levantamento inicial dos Acórdãos e dos editais objeto do julgamento, deve-se escolher os tópicos dos julgados considerados relevantes para o desenvolvimento do trabalho. Uma vez que os dados sejam selecionados e preparados para a correta identificação das previsões editalícias em desconformidade com a jurisprudência, um classificador textual poderá ser modelado de modo a identificar as eventuais irregularidades e/ou impropriedades.

REFERÊNCIAS

CASTRO, Cristiano Leite de; BRAGA, Antônio Pádua. Aprendizado supervisionado com conjuntos de dados desbalanceados. Sba Controle & Automação, Campinas , v. 22, n. 5, p. 441-466, out. 2011 . Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-17592011000500002&lng=pt&nrm=iso>. Acesso em: 09 mar. 2020. https://doi.org/10.1590/S0103-17592011000500002

BRASIL. Manual Técnico do Orçamento. 2020. Disponível em <https://www1.siop.planejamento.gov.br/mto/doku.php/mto2020>. Acesso em: 09 mar.

CORTES, Corinna; VAPNIK, Vladimir. Support-vector networks. Machine learning, 20(3):273–297, 1995

LEWIS, David D. Naive (bayes) at forty: The independence assumption in information Retrieval, 1998.

BREIMAN, Leo. Random forests. Technical report, Technical Report 567, Department of Statistics, UC Berkeley, 1999.

ZAKI, Mohammed J.; Jr., Wagner Meira. Data Mining and Analysis. Cambridge University Press, 2014.

ZHANG, J., YANG, Y. "Robustness of regularized linear classification methods in text categorization". In: Annual ACM Conference on Research and Development in Information Retrieval, pp. 190-197, Toronto, 2003.

CHAPMAN, Pete et al. CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, v. 16, 2000.

BRASIL. Portaria Interministerial STN/SOF Nº 163, de 4 de maio de 2001. Dispõe sobre normas gerais de consolidação das Contas Públicas no âmbito da União, Estados, Distrito Federal e Municípios, e dá outras providências, 2001.

BRASIL. Decreto 1094, de 23 de março de 1994.

BRASIL. Projeto de Lei 1292, de 30 de novembro de 1995. Altera a Lei 8.666, de 21 de junho de 1993, que regulamenta o art. 37, inciso XXI, da Constituição Federal, institui normas para licitações e contratos da Administração Pública e dá outras providências. Disponível em < https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=16526>. Acesso em: 10 mar. 2020.

PATI, Camila. Estas profissões podem acabar até 2030 (ao menos para humanos). Exame, São Paulo, 21 dez. 2017. Disponível em <https://exame.abril.com.br/carreira/estas-profissoes-podem-acabar-ate-2030-ao-menos-para-os-humanos/>. Acesso em: 10 mar. 2020.

SEBASTIANI, Fabrizio. Machine Learning in Automated Text Categorization. ACM Computing Surveys. 2002.