UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS GRADUAÇÃO PROFISSIONAL EM ENGENHARIA DE PRODUÇÃO

EDILSON PEREIRA MARINS

APLICAÇÃO DE CLUSTERING PARA APOIO À ANÁLISE DE RISCO DE DESEQUILÍBRIO ECONÔMICO-FINANCEIRO NAS ATAS DE REGISTRO DE PREÇO DA FUFMT

SÃO CARLOS – SP
2026

EDILSON PEREIRA MARINS

APLICAÇÃO DE CLUSTERING PARA APOIO À ANÁLISE DE RISCO DE DESEQUILÍBRIO ECONÔMICO-FINANCEIRO NAS ATAS DE REGISTRO DE PREÇO DA FUFMT

Dissertação apresentada ao Programa de Pós-Graduação Profissional em Engenharia de Produção da Universidade Federal de São Carlos para obtenção do título de Mestre em Engenharia de Produção.

Orientador: Prof. Dr. Roberto Fernandes Tavares Neto

São Carlos – SP
2026

EDILSON PEREIRA MARINS

Aplicação de *Clustering* Para Apoio à Análise de Risco de Desequilíbrio Econômico-Financeiro nas Atas de Registro de Preço da FUFMT

Dissertação apresentada ao Programa de Pós-Graduação Profissional em Engenharia de Produção da Universidade Federal de São Carlos para obtenção do título de Mestre em Engenharia de Produção.

BANCA EXAMINADORA

gov.br
Documento assinado digitalmente
ROBERTO FERNANDES TAVARES NETO
Data: 28/04/2026 13:58:03-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Roberto Fernandes Tavares Neto (Orientador)

Profa. Dra. Cendyi Aparecida Paes de Barros do Prado (UFMT)

Prof. Dr. Fabio Molina da Silva (UFSCar)

São Carlos, 23 de março de 2026

RESUMO

O presente trabalho investiga o processo de reequilíbrio econômico-financeiro nas Atas de Registro de Preço no âmbito da Universidade Federal de Mato Grosso, propondo a aplicação de técnicas de análises estatísticas e aprendizado de máquina para possibilitar a otimização do processo decisório e de gestão das contratações públicas. O estudo desenvolve uma ferramenta tecnológica, por meio de uso de linguagem Python, com capacidade de realizar a classificação e agrupamento de documentos licitatórios da Instituição, tais como editais, atas e ordens de fornecimento, com base na similaridade textual dos mesmos, visando identificar padrões relacionados à ocorrência de desequilíbrio econômico-financeiro nas contratações. A pesquisa foi desenvolvida adotando-se uma abordagem quantitativa, por meio de uso de modelagem e simulação, assim como implementação de algoritmos de *clustering*, com ênfase no *K-Means*, permitindo a realização de análise de dados não estruturados. Os resultados da pesquisa demonstram a viabilidade de aplicação de técnicas de inteligência artificial (IA) como suporte no processo decisório na gestão pública, permitindo a identificação de contratações com maior probabilidade de ocorrência de instabilidades contratuais, fornecendo subsídios para um monitoramento de forma preventiva nos processos licitatórios. A proposta da pesquisa contribui para um aprimoramento da eficiência administrativa, da transparência e da sustentabilidade na utilização de recursos públicos, fornecendo um *framework* replicável em outras instituições públicas.

Palavras-chave: Reequilíbrio econômico-financeiro. *Clustering*. Administração Pública.

ABSTRACT

This study investigates the economic-financial rebalancing process in Price Registration Agreements within the Federal University of Mato Grosso (UFMT), proposing the application of statistical analysis and machine learning techniques to optimize decision-making and the management of public procurement. The research develops a technological tool, implemented in Python, capable of classifying and clustering the institution's bidding documents, such as public notices, agreements, and supply orders, based on their textual similarity, aiming to identify patterns related to the occurrence of economic-financial imbalance in contracts. The study adopts a quantitative approach through modeling and simulation, as well as the implementation of clustering algorithms, with emphasis on K-means, enabling the analysis of unstructured data. The results demonstrate the feasibility of applying artificial intelligence (AI) techniques as a decision-support tool in public management, allowing the identification of contracts with a higher likelihood of instability and providing a foundation for preventive monitoring of procurement processes. The research proposal contributes to improving administrative efficiency, transparency, and sustainability in the use of public resources, offering a framework that can be replicated in other public institutions.

Keywords: Economic-financial rebalancing. Clustering. Public administration.

LISTA DE SIGLAS

ARP	Ata de Registro de Preços
BERT	*Bidirectional Encoder Representations from Transformers*
BoW	*Bag of Words*
CBOW	*Continuous Bag-of-Words*
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*
EPP	Empresa de Pequeno Porte
GloVe	*Global Vectors for Word Representation*
IA	Inteligência Artificial
IDF	Frequência Inversa do Documento
IFES	Instituição Federal de Ensino Superior
LOA	Lei Orçamentária Anual
ME	Micro Empresa
NE	Nota de Empenho
NLTK	*Natural Language Toolkit*
OCR	Reconhecimento Óptico de Caracteres
OF	Ordem de Fornecimento
PCA	Análise de Componentes Principais
PIB	Produto Interno Bruto
PLN	Processamento de Linguagem Natural
SEI	Sistema Eletrônico de Informações
SRP	Sistema de Registro de Preços
TCU	Tribunal de Contas da União
TF	Termo de Frequência
TF-IDF	*Term Frequency-Inverse Document Frequency*
UFMT	Universidade Federal de Mato Grosso

LISTA DE FIGURAS

Figura 1 - Etapas do Processo de Licitação 17
Figura 2 - Etapas do Registro de Preços 19
Figura 3 - Interligação Temática da Inteligência Artificial (IA) 21
Figura 4 - Fluxograma de Execução do Algoritmo *K-Means* 33
Figura 5 - Visualização Gráfica de SSE com Curva Clara 36
Figura 6 - Testes iniciais, com 100 arquivos, 12m48s 56
Figura 7 - Testes iniciais, com 2500 arquivos, 2h25m47s 56
Figura 8 - Extração de texto 60
Figura 9 - Pré-processamento de texto 60
Figura 10 - Remoção de stopwords 61
Figura 11 - Aplicação do *Word2Vec* 65
Figura 12 - Implementação do *clustering* 68
Figura 13 - Exemplo de visualização gráfica em espaço 2D 69
Figura 14 - Visualização da análise com k = 3 72
Figura 15 - Visualização do *cluster* “0”, com k = 3 73
Figura 16 - Visualização da análise com k = 4 74
Figura 17 - Visualização da análise com k = 5 74
Figura 18 - Visualização da análise com k = 6 75

LISTA DE QUADROS

Quadro 1 - Análise Comparativa de Modelos de *Word Embedding* 39
Quadro 2 - Comparação entre Representações Esparsas e Densas.....39
Quadro 3 - Atributos da Classificação de Processo de Reequilíbrio 57
Quadro 4 - Análise Comparativa dos algoritmos *K-Means* e DBSCAN.....66
Quadro 5 - Comparativo de Agrupamento com *k* Variável 70
Quadro 6 - Resumo da Composição dos *Clusters* (*k* = 3).....72
Quadro 7 - Resumo da Composição dos *Clusters* (*k* = 6).....75
Quadro 8 - Análise da Composição dos *Clusters* (*k* = 5) e Percentual de Concentração de Reequilíbrio 77

SUMÁRIO

1 INTRODUÇÃO 10
1.1 OBJETIVOS 12
1.1.1 Objetivo geral 12
1.1.2 Objetivos específicos 13
1.2 JUSTIFICATIVA 13
2 REFERENCIAL TEÓRICO 16
2.1 LICITAÇÃO 16
2.1.1 Registro de Preços 17
2.2 INTELIGÊNCIA ARTIFICIAL E *MACHINE LEARNING* 19
2.2.1 *Clustering* 22
2.2.2 Algoritmo *K-Means* 30
2.2.2.1 Aprimorando a inicialização com K-Means++ 34
2.2.2.2 Método do Cotovelo (Elbow Method) 35
2.2.3 Classificação Textual 37
2.2.3.1 Inteligência de dados textuais em aquisições 40
2.2.3.1.1 O uso do processamento de linguagem natural para análise documental . 40
3 METODOLOGIA 42
3.1 FUNDAMENTAÇÃO DA ABORDAGEM METODOLÓGICA 42
3.2 JUSTIFICATIVA DO MÉTODO: MODELAGEM E SIMULAÇÃO 44
4 PROPOSTA DE *FRAMEWORK* PARA CLASSIFICAÇÃO DE LICITAÇÕES 47
4.1 COLETA E PREPARAÇÃO DOS DADOS 47
4.1.1 Coleta de dados 47
4.1.2 Pré-processamento dos dados 48
4.2 REPRESENTAÇÃO DOS DADOS TEXTUAIS 49
4.3 APLICAÇÃO DO ALGORITMO *K-MEANS* 50
4.3.1 Definição do Número de *Clusters* 50
4.4 INTERPRETAÇÃO DOS *CLUSTERS* E APLICAÇÃO DE MODELOS PREDITIVOS 51
4.4.1 Análise dos *Clusters* 51
4.4.2 Uso de Modelos Preditivos 51
4.5 DAS LIMITAÇÕES NO DESENVOLVIMENTO DA PESQUISA 52

5 RESULTADOS E DISCUSSÕES 54
5.1 DEFINIÇÃO E DETALHAMENTO DOS DADOS UTILIZADOS 54
5.1.1 Dos atributos para agrupamento dos processos de reequilíbrio analisados entre 2020 e 2022 56
5.2 CONVERSÃO DE PDFS NÃO ESTRUTURADOS EM TEXTO PROCESSÁVEL 59
5.2.1 Transformação de texto em representações quantitativas 61
5.2.2 O modelo *Word2Vec* e as arquiteturas CBOW e *Skip-Gram* 63
5.3 APRENDIZADO NÃO SUPERVISIONADO PARA A DESCOBERTA DE PADRÕES EM DOCUMENTOS DE AQUISIÇÃO 65
5.3.1 Princípios do Agrupamento Baseado em Densidade e Baseado em Centróides 65
5.3.2 Implementação com código Python e *Scikit-learn* 67
5.4 ANÁLISE EMPÍRICA E INTERPRETAÇÃO DOS *CLUSTERS* DE AQUISIÇÕES 69
5.4.1 Definição do número ótimo de *clusters* com aplicação do método do cotovelo 69
5.4.2 Avaliação de resultados com números variados de *clusters* 71
5.4.2.1 Dos resultados quando k = 3 71
5.4.2.2 Dos resultados quando k = 4, k = 5 e k = 6 73
5.4.3 Análise dos resultados e potencialidade preditiva do modelo 76
5.5 IMPLICAÇÕES ESTRATÉGICAS E APLICAÇÕES PARA A GESTÃO DO SETOR PÚBLICO 78
5.5.1 Inteligência acionável na gestão de aquisições 79
5.5.2 Sistema preditivo para a mitigação do desequilíbrio econômico-financeiro 79
6 CONSIDERAÇÕES FINAIS 81
REFERÊNCIAS 83
APÊNDICE A: PRODUTO TECNOLÓGICO 89

1 INTRODUÇÃO

Historicamente, a universidade tem sido concebida como um espaço de produção e difusão do conhecimento, cuja missão social transcende a mera formação de profissionais e abrange a promoção da inclusão, da pesquisa e da extensão. Conforme demonstrado por Caetano e Campos (2019), a autonomia das universidades federais na execução de suas receitas próprias é elemento crucial para o fortalecimento de sua função social, permitindo a adaptação de seus processos internos às demandas específicas da comunidade acadêmica e da sociedade.

Desde a promulgação da Constituição Federal de 1988, o orçamento destinado à educação passou por intensas transformações. A Lei Orçamentária Anual (LOA), que define as diretrizes e os limites dos gastos públicos a cada exercício, e o orçamento discricionário, composto por recursos de natureza variável, destinados a atender demandas imediatas e estratégias de gestão, são institutos fundamentais para a transparência e o planejamento orçamentário no setor.

Entretanto, os investimentos nas universidades federais vêm sendo progressivamente reduzidos, inclusive com a promulgação da Emenda Constitucional nº 95/2016, que impôs um teto aos gastos públicos, reduzindo significativamente o orçamento das Instituições Federais de Ensino Superior (IFES) e afetando diretamente a manutenção e a expansão das atividades acadêmicas e de pesquisa, conforme evidenciado por Inácio *et al*. (2021).

Nesse cenário, a autonomia administrativa e financeira, historicamente conferida às universidades para gerir seus recursos próprios, vem sendo constantemente desafiada pelas restrições orçamentárias impostas pelo contexto macroeconômico e pelas políticas públicas vigentes. Assim, a compreensão detalhada dos mecanismos de alocação dos recursos, incluindo o planejamento orçamentário definido na LOA e o uso do orçamento discricionário, revela a necessidade de estratégias inovadoras para garantir não apenas a manutenção, mas a ampliação da função social dessas instituições.

Diante deste contexto de austeridade financeira, no que tange às compras públicas, a nova Lei nº 14.133/2021 trouxe importantes inovações aos procedimentos licitatórios, definindo diversas modalidades de contratação. Entre elas, o pregão eletrônico destaca-se por sua agilidade, transparência e competitividade, facilitando a

aquisição de bens e serviços com economia e eficiência administrativa, conforme demonstrado por Moreira *et al*. (2024). Contudo, apesar dos avanços normativos, a execução dos processos licitatórios ainda enfrenta desafios práticos, desde o planejamento das compras, a percepção de diferenciação de preços em diferentes modalidades de compra, como demonstrado por Nunes e Dantas (2012), e na manutenção do equilíbrio econômico contratual durante a vigência das atas de registro de preços.

O Sistema de Registro de Preços (SRP), procedimento auxiliar previsto na Lei n.º 14.133/2021 e regulamentado pelo Decreto n.º 11.462/2023, possibilita a padronização e a sistematização das aquisições públicas, permitindo a realização de compras com base em um procedimento previamente estabelecido. Estudos recentes apontam que, embora o SRP apresente vantagens como a economia de escala e a redução de custos operacionais, ele também sofre limitações quanto à flexibilidade no planejamento de longo prazo, à diferenciação de preços e ao risco de desequilíbrio econômico durante a vigência das atas, como destacado por Machado *et al*. (2018) e Nunes e Dantas (2012).

A presente pesquisa tem como escopo a análise de padrões em processos licitatórios no âmbito da Universidade Federal de Mato Grosso (UFMT), com foco em contratações realizadas por meio do Sistema de Registro de Preços. Nesse contexto, busca-se responder à seguinte questão de pesquisa: como identificar padrões em processos licitatórios, por meio de técnicas de *clustering*, associados à ocorrência de desequilíbrio econômico-financeiro nas atas de registro de preços?

A investigação parte do pressuposto de que os documentos que compõem esses processos contêm informações relevantes capazes de revelar similaridades estruturais e contextuais, possibilitando a identificação de agrupamentos com características comuns. Dessa forma, a análise proposta contribui para a compreensão de padrões relacionados ao desequilíbrio econômico-financeiro, com potencial de subsidiar a tomada de decisão na gestão pública.

Assim, o objetivo geral desta pesquisa consiste no desenvolvimento de uma ferramenta baseada em técnicas de *machine learning* para análise de documentos licitatórios já realizados. Especificamente, a abordagem proposta envolve a utilização de métodos de clusterização (agrupamento) textual, sendo esta última caracterizada como uma técnica de aprendizado não supervisionado que permite agrupar

documentos com base em sua similaridade semântica. A aplicação de *clustering* justifica-se pela natureza não estruturada dos dados analisados e pela ausência de categorias previamente definidas, possibilitando a identificação de padrões latentes nos documentos licitatórios que não seriam facilmente detectáveis por meio de análises tradicionais. A partir desses agrupamentos, torna-se possível analisar a concentração de ocorrências de reequilíbrio econômico-financeiro em conjuntos de documentos com características semelhantes, permitindo inferir padrões associados ao risco contratual. Dessa forma, a ferramenta proposta visa auxiliar na predição de ocorrências de reequilíbrio nas Atas de Registro de Preços, oferecendo aos gestores públicos um instrumento analítico capaz de identificar licitações com maior probabilidade de instabilidade econômico-financeira ao longo de sua vigência, contribuindo para a mitigação de riscos e o aprimoramento do processo decisório. Diferentemente de abordagens supervisionadas, que exigem rotulação prévia dos dados, a clusterização permite uma análise exploratória mais flexível, especialmente adequada ao contexto de documentos administrativos heterogêneos.

A aplicação de técnicas de aprendizado de máquina no processo decisório, especialmente aquelas baseadas em métodos de agrupamento de dados (*clustering*), enfrenta desafios inerentes quando aplicada a conjuntos de dados de alta complexidade, como aqueles derivados de documentos textuais não estruturados. Esses desafios são cruciais para a compreensão, interpretação e aplicação eficaz da técnica, especialmente em cenários onde a decisão baseada em dados assume um papel preponderante (Melo *et al*., 2022). Com o emprego dessa solução tecnológica, será possível aprimorar o planejamento das compras públicas, contribuindo para uma alocação mais eficiente dos recursos e para o fortalecimento da autonomia das instituições de ensino superior.

1.1 OBJETIVOS

1.1.1 Objetivo geral

Desenvolver ferramenta tecnológica que possibilite a identificação de processos de licitação com maior probabilidade de ocorrência de desequilíbrio

econômico-financeiro nas Atas de Registro de Preço da FUFMT, possibilitando aos gestores a mitigação de riscos.

O produto tecnológico a ser desenvolvido terá como foco a melhoria do acompanhamento das licitações, permitindo sinalizar determinados certames para que o monitoramento durante sua vigência seja mais diligente. O produto pode ainda ser replicado em outras instituições, tendo em vista a similaridade de processos de reequilíbrio existentes.

1.1.2 Objetivos específicos

- Analisar de modo geral o fluxo dos processos de reequilíbrio na FUFMT;
- Identificar os principais atributos existentes em processos de reequilíbrio;
- Analisar os artefatos de licitações realizadas pela UFMT, de modo a obter um banco de dados de treinamento para desenvolvimento da ferramenta tecnológica; e
- Clusterizar os artefatos analisados com base nas similaridades entre os mesmos.

1.2 JUSTIFICATIVA

No contexto de restrição orçamentária enfrentado pelas Instituições Federais de Ensino Superior, o suporte para uma aplicação eficiente dos recursos torna-se imperativa, uma vez que a gestão eficiente dos orçamentos é fundamental para assegurar que a instituição possa manter sua missão social de forma sustentável.

A presente pesquisa se propõe a desenvolver uma abordagem analítica voltada à identificação de padrões associados à ocorrência de desequilíbrio econômico-financeiro durante a vigência das Atas de Registro de Preços, por meio de técnicas de agrupamento aplicadas a documentos licitatórios. A utilização dessa abordagem permite não apenas identificar processos com maior probabilidade de ocorrência de desequilíbrio, mas também subsidiar decisões administrativas mais direcionadas e preventivas. Nesse sentido, os resultados podem orientar, por exemplo, o aumento do nível de monitoramento de determinados contratos, a priorização de auditorias e

monitoramento diligente em processos agrupados como de maior risco, o reforço no controle orçamentário durante a execução contratual, bem como o aprimoramento do planejamento das aquisições futuras. Dessa forma, a proposta contribui para uma atuação mais proativa da Administração Pública, promovendo maior eficiência na gestão dos recursos públicos.

O desenvolvimento de uma ferramenta que auxilie na análise preditiva dos processos licitatórios, identificando de forma antecipada potenciais desequilíbrios e permitindo a tomada de decisões que mitiguem riscos e promovam a eficiência se mostra fundamental. A proposta de utilizar técnicas de *machine learning* e análise de dados para agrupamento de documentos licitatórios insere-se em um campo emergente de aplicação de inteligência artificial na gestão pública. Embora a literatura recente apresente avanços na utilização de técnicas de análise de dados em compras públicas, tais aplicações ainda se concentram, majoritariamente, em tarefas como detecção de fraudes, análise de preços e apoio à tomada de decisão em licitações.

Observa-se, contudo, uma lacuna no que se refere à utilização de técnicas de processamento de linguagem natural e aprendizado não supervisionado para análise de documentos licitatórios com foco na identificação de padrões associados ao desequilíbrio econômico-financeiro, especialmente no contexto do Sistema de Registro de Preços.

Nesse sentido, a presente pesquisa contribui ao propor uma abordagem metodológica que integra técnicas de clusterização textual e análise preditiva aplicada a dados reais de uma instituição pública, ampliando o escopo de utilização de métodos de inteligência artificial no apoio à gestão de contratos administrativos, oferecendo subsídio tecnológico essencial para os gestores públicos. Ademais, a aplicação dessa abordagem em dados administrativos reais reforça sua relevância prática, aproximando a pesquisa acadêmica das demandas concretas da gestão pública. A ferramenta proposta não apenas refina a análise dos dados dos processos licitatórios, mas também contribui para a transparência e a *accountability* na aplicação dos recursos ao permitir a geração de evidências analíticas que fundamentam a tomada de decisão.

Ao agrupar documentos com base em padrões identificáveis, a ferramenta possibilita a rastreabilidade dos critérios utilizados na análise dos processos, reduzindo a subjetividade das decisões administrativas. Além disso, ao indicar

contratações com maior probabilidade de ocorrência de desequilíbrio econômico-financeiro, fornece subsídios objetivos para a priorização de ações de controle, como auditorias, inspeções e monitoramento contratual, fortalecendo os mecanismos de governança e controle interno.

Com isso, espera-se possibilitar um monitoramento mais diligente em relação às licitações sinalizadas com maior potencial de ocorrências de desequilíbrio econômico-financeiro, o que é crucial num cenário de restrição orçamentária. Dessa forma, o trabalho propõe uma abordagem que alia inovação tecnológica à eficiência administrativa, contribuindo para a tomada de decisão baseada em dados (*data-driven decision making*), alinhando-se às diretrizes contemporâneas de modernização da administração pública, atendendo à necessidade de políticas públicas mais precisas e sustentáveis.

2 REFERENCIAL TEÓRICO

2.1 LICITAÇÃO

A licitação é o procedimento administrativo por meio do qual o poder público seleciona, de forma isonômica e competitiva, a proposta que melhor atenda aos interesses coletivos na contratação de bens, serviços ou obras. Seu objetivo é garantir a eficiência, a transparência e a legalidade dos gastos públicos, assegurando que os recursos do erário sejam aplicados de maneira otimizada. No Brasil, os fundamentos legais da licitação encontram respaldo na Constituição Federal de 1988, que consagra os princípios da legalidade, da impessoalidade, da moralidade, da publicidade e da eficiência na administração pública, e na Lei nº 14.133/2021, que substitui a Lei nº 8.666/1993 e traz inovações quanto à modernização dos processos licitatórios. Dentre as modalidades licitatórias, o pregão eletrônico se destaca em função de promover maior celeridade e ampliar a competitividade, muito em função da redução de prazos e de exigências documentais (Freitas; Maldonado, 2013), alinhando-se aos princípios constitucionais e à necessidade de racionalização dos gastos públicos.

A Nova Lei de Licitações (Lei n.º 14.133/2021) organiza o processo de licitação em etapas internas e externas. O artigo 17 da referida lei estabelece, de forma resumida, a ordem cronológica do processo, enquanto os artigos 18 a 71 detalham cada uma das mesmas.

Fase Interna (Preparatória)

1. Planejamento da Contratação
    a. Identificação da necessidade;
    b. Elaboração dos estudos técnicos preliminares (ETP);
    c. Análise de riscos;
    d. Elaboração do Termo de Referência ou Projeto Básico (definição da modalidade, orçamentos e análise de viabilidade).
2. Autorização e Designação do Agente
    a. Autorização da Autoridade Competente;
    b. Designação da comissão responsável;
    c. Elaboração do edital do certame;
    d. Análise jurídica.

Fase Externa (Procedimental)

3. Publicação do edital
    a. Divulgação da licitação nos portais oficiais;
    b. Prazo para impugnações e esclarecimentos.
4. Recebimento de propostas e disputa
    a. Envio de propostas pelos interessados;
    b. Execução da fase disputa (lances).
5. Julgamento
    a. Análise das propostas em relação ao exigido no edital.
6. Habilitação
    a. Análise das documentações dos licitantes em relação ao exigido no edital.
7. Etapa recursal
    a. Interposição e julgamento de recursos.
8. Adjudicação
    a. Declaração, por parte da Autoridade Competente, do vencedor da licitação.
9. Homologação
    a. Validação, por parte da Autoridade Competente, do procedimento licitatório como um todo e finalização do mesmo.

Figura 1 - Etapas do Processo de Licitação

Fonte: elaborado pelo autor (2026)

2.1.1 Registro de Preços

Adentrando mais ao *rol* de modalidades de licitação e procedimentos auxiliares, o sistema de registro de preços (SRP) configura-se como uma importante ferramenta

utilizada para padronizar a aquisição de bens e a contratação de serviços, permitindo que a administração pública contrate, ao longo do tempo, o fornecimento de determinados itens com preços e condições previamente estabelecidos. O SRP está previsto como procedimento auxiliar no artigo 78 da Lei nº 14.133/2021 e foi posteriormente regulamentado pelo Decreto nº 11.462/2023, tendo como objetivos principais garantir a economicidade, possibilitar uma melhor previsão dos custos e facilitar a execução dos contratos, sobretudo em contextos onde as demandas são recorrentes e com alto nível de padronização (Santos; Ziger; Michelon, 2022).

As etapas do Registro de Preços, conforme estabelecem a Lei n.º 14.133/2021 e o Decreto n.º 11.462/2023, são estabelecidas em sequência ao procedimento de licitação. Ou seja, concluídas as etapas do processo de licitação, até a etapa de homologação pela Autoridade Competente, segue-se então com:

1. Assinatura da Ata de Registro de Preços
    a. Finalizada a licitação, os licitantes vencedores são convocados para a formalização das ARPs, conforme itens/lotes vencidos.
2. Levantamento de demanda
    a. A Instituição define os quantitativos a serem adquiridos, conforme planejamento interno.
3. Indicação de disponibilidade orçamentária
    a. Definidas as quantidades a serem adquiridas, a Instituição deve realizar a indicação de disponibilidade de orçamento para efetivar a contratação ou aquisição.
4. Emissão do empenho
    a. A Instituição então emite a Nota de Empenho (NE), por meio do qual vincula a contratação ou aquisição ao orçamento disponibilizado. Desta forma, os valores ficam “reservados” para aquela contratação.
5. Elaboração da Ordem de Fornecimento ou Contrato
    a. A etapa final da formalização da contratação se dá com a emissão e assinatura do contrato ou da Ordem de Fornecimento (OF).

Figura 2 - Etapas do Registro de Preços

Fonte: elaborado pelo autor (2026)

A praticidade temporal do Registro de Preços se evidencia de forma marcante quando avaliamos a Figura 2 acima, uma vez que tendo sido previamente realizada a licitação e assinada a ARP, apenas as etapas 3, 4 e 5 se repetirão quando for necessária a contratação ou aquisição do objeto registrado.

Entretanto, embora o SRP ofereça vantagens notáveis como a redução de custos operacionais, padronização dos processos e celeridade, estudos recentes apontam que ele também apresenta desafios relacionados à flexibilidade no planejamento a longo prazo e à manutenção do equilíbrio econômico durante a vigência das atas (Gonçalves *et al*., 2018; Nunes; Dantas, 2012).

2.2 INTELIGÊNCIA ARTIFICIAL E *MACHINE LEARNING*

A Inteligência Artificial (IA) é uma área da ciência da computação dedicada ao desenvolvimento de sistemas capazes de executar tarefas que, tradicionalmente, requerem inteligência humana, tais como reconhecimento de padrões, tomada de decisão e aprendizado. No interior desse campo, o aprendizado de máquina (*machine learning*) constitui um subcampo fundamental, baseado no desenvolvimento de modelos computacionais capazes de aprender padrões a partir de dados e generalizar esse conhecimento para novos casos. Autores como Russell e Norvig (2021) e Goodfellow *et al*. (2016) demonstram que o *machine learning* tem revolucionado diversos setores ao possibilitar a automação e a análise preditiva em larga escala. Diferentemente de abordagens tradicionais, nas quais as regras são explicitamente programadas, o *machine learning* utiliza métodos estatísticos e algoritmos de otimização para ajustar modelos a partir de exemplos, buscando minimizar erros ou maximizar critérios de desempenho previamente definidos. Esses métodos podem ser

classificados, de forma geral, em abordagens supervisionadas, nas quais o modelo aprende a partir de dados rotulados, e não supervisionadas, que visam identificar estruturas e padrões intrínsecos em dados sem rótulos, como no caso das técnicas de *clustering*.

No contexto desta pesquisa, destaca-se o uso de aprendizado não supervisionado, especialmente por meio de algoritmos de agrupamento, como forma de identificar similaridades estruturais entre documentos licitatórios e extrair padrões relevantes para a análise do desequilíbrio econômico-financeiro. Do ponto de vista formal, o aprendizado de máquina pode ser compreendido como o processo de estimar uma função a partir de dados, de modo a capturar relações subjacentes entre variáveis e permitir previsões ou inferências sobre novos dados.

Observa-se, no entanto, que a evolução da Inteligência Artificial (AI) aconteceu de forma não linear, sendo marcada por ciclos notáveis de avanços e outros de retrocessos. Na visão proposta por Toosi *et al*. (2021), estes ciclos são denominados historicamente como “verões” e “invernos”, respectivamente, marcados por ondas de elevado otimismo tecnológico, seguidas de significativa frustração em função de resultados práticos.

A recente expansão da inteligência artificial está associada a um conjunto de avanços tecnológicos e estruturais que ampliaram significativamente sua capacidade de aplicação em diferentes domínios. Dentre os principais fatores, destacam-se: (i) o aumento da capacidade computacional, viabilizado por arquiteturas de processamento paralelo, como GPUs; (ii) a ampla disponibilidade de dados em larga escala, incluindo bases governamentais digitais; e (iii) o desenvolvimento de modelos mais sofisticados, especialmente no campo do aprendizado profundo (*deep learning*).

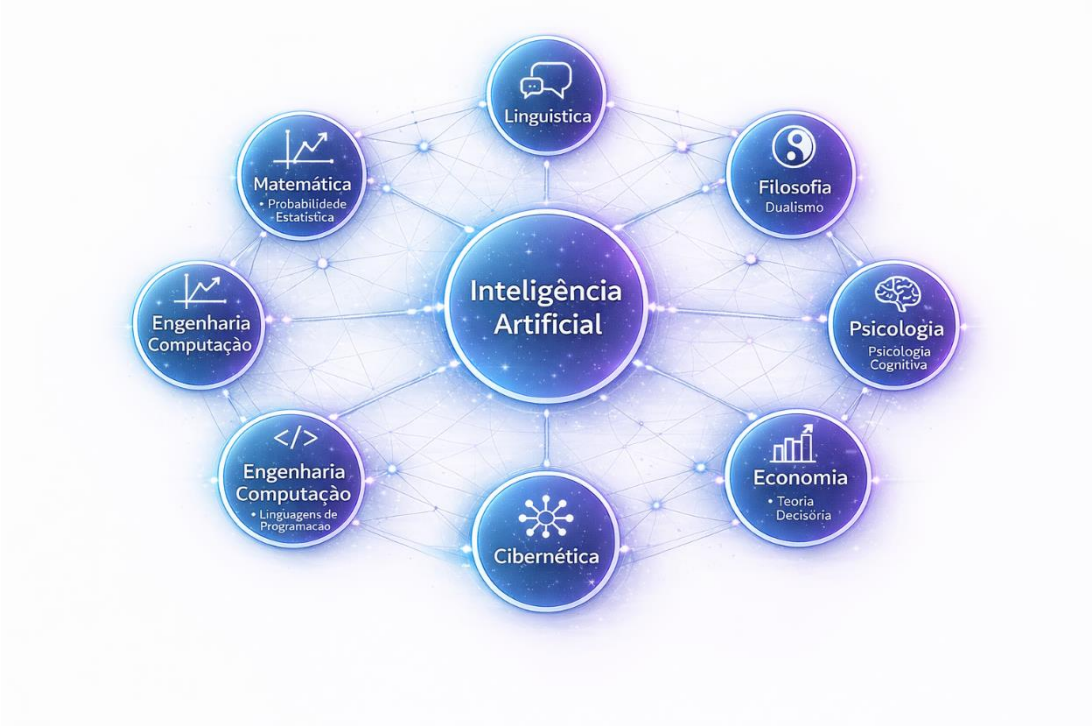
Esses elementos combinados possibilitam o treinamento de modelos mais complexos e a extração de padrões em conjuntos de dados de alta dimensionalidade e natureza não estruturada, como textos. Nesse contexto, técnicas de aprendizado de máquina tornam-se particularmente relevantes para a análise de documentos administrativos, permitindo identificar relações e regularidades que não são facilmente capturadas por abordagens tradicionais baseadas em regras explícitas.

No setor público, a crescente digitalização dos processos administrativos, incluindo a produção e armazenamento de documentos licitatórios, cria um ambiente propício para a aplicação dessas técnicas. Assim, a utilização de métodos de

inteligência artificial para análise desses dados representa uma oportunidade de aprimorar a capacidade analítica da administração pública, especialmente no que se refere à identificação de padrões e apoio à tomada de decisão. Essa capacidade é particularmente relevante em contextos nos quais a tomada de decisão depende da análise de grandes volumes de informação textual, como no caso das compras públicas.

Toosi *et al*. (2021) evidenciam uma ampla aplicabilidade da inteligência artificial em diversos setores, destacando toda a potencialidade de superar paradigmas computacionais, além de possibilitar ganhos significativos quanto a eficiência, predição e escalabilidade. Os autores mostram que o desenvolvimento recente não se deve puramente pelo aumento de performance computacional, envolvendo ainda a evolução de arquiteturas de algoritmos, incorporação de grandes volumes de dados e lapidação de modelos de *deep learning*.

Figura 3 - Interligação Temática da Inteligência Artificial (IA)



Fonte: adaptado de Toosi *et al*. (2021)

No contexto desta pesquisa, o emprego de técnicas de *machine learning* possibilita a identificação de padrões em documentos licitatórios, contribuindo para a predição de possíveis ocorrências de desequilíbrios econômico-financeiros durante a vigência das atas de registro de preços, surgindo como uma aplicação estratégica

para a Administração Pública, alinhada aos princípios da eficiência e economicidade, os quais são previstos na Lei n.º 14.133/2021. Russell e Norvig (2021) argumentam que a capacidade de aprendizado de máquina com base em dados históricos, adaptando-se a novos padrões, mostra-se como um novo paradigma no processo decisório governamental.

2.2.1 *Clustering*

O *clustering*, ou agrupamento de dados, é uma técnica de aprendizado não supervisionado cujo objetivo é organizar um conjunto de observações em grupos (*clusters*) de forma que elementos pertencentes ao mesmo grupo apresentem alta similaridade entre si, enquanto elementos de grupos distintos sejam significativamente diferentes (Han; Kamber; Pei, 2011).

Diferentemente das abordagens supervisionadas, nas quais os dados de entrada estão previamente rotulados, o *clustering* busca identificar estruturas latentes nos dados sem a existência de categorias predefinidas, permitindo a descoberta de padrões intrínsecos a partir das próprias características dos dados (Jain, 2010).

A finalidade do *clustering* está associada à análise exploratória de dados, sendo amplamente utilizado para segmentação, identificação de padrões, detecção de anomalias e redução de complexidade em grandes conjuntos de dados. No contexto de dados textuais, essa técnica permite agrupar documentos com base em similaridades semânticas, possibilitando a identificação de temas, estruturas ou comportamentos recorrentes.

No âmbito desta pesquisa, o *clustering* é utilizado como ferramenta para identificar padrões estruturais em documentos licitatórios, permitindo analisar a concentração de ocorrências de desequilíbrio econômico-financeiro em grupos de processos com características semelhantes. Dessa forma, a técnica contribui para a extração de conhecimento a partir de dados não estruturados, apoiando a tomada de decisão na gestão pública.

A utilização de técnicas de *clustering* extrapola os limites do ramo computacional e tem se consolidado com um mecanismo analítico fundamental em diversos campos. Diversos estudos apontam o uso da clusterização como ferramenta essencial no desenvolvimento de pesquisas nos mais variados setores. Como

destacam Ventorim, Luchi e Varejão (2020), o *clustering* pode ser utilizado na administração pública para identificar padrões de comportamento em processos de contratação, de forma a avaliar perfis de fornecedores e detectar inconformidades na execução contratual. Segundo os autores, o agrupamento de dados possibilita a identificação de correlações subjetivas, contribuindo para o processo decisório de maneira mais robusta e transparente.

No ramo financeiro, técnicas de clusterização são amplamente utilizadas na segmentação de clientes, em detecção de fraudes e na análise de risco de crédito. Jain (2010) e Aggarwal (2018) demonstram em seus estudos que algoritmos de agrupamento permitem detectar e agrupar clientes por meio de padrões de consumo semelhantes. Em processos de auditoria, técnicas de *clustering* têm sido amplamente utilizadas na identificação de padrões atípicos em grandes volumes de dados, contribuindo para a detecção de transações suspeitas e o apoio a atividades de controle e prevenção de irregularidades (Chandola; Banerjee; Kumar, 2009). Essas abordagens permitem identificar comportamentos fora do padrão esperado, sendo particularmente úteis em contextos nos quais não há rótulos previamente definidos para caracterizar fraudes ou inconsistências.

O uso de técnicas de agrupamento também é recorrente na área da saúde e nas ciências biológicas. Na bioinformática, por exemplo, o *clustering* é empregado para a identificação de padrões em dados genéticos e expressão gênica, auxiliando na descoberta de estruturas biológicas relevantes (Jain, 2010). Na área médica, essas técnicas são utilizadas na análise de imagens clínicas e na segmentação de regiões com características semelhantes, contribuindo para diagnósticos mais precisos (Larranaga *et al*., 2006). Além disso, em estudos epidemiológicos, o *clustering* é aplicado para o agrupamento de populações e regiões com padrões semelhantes de ocorrência de doenças, apoiando a formulação de políticas públicas em saúde (Han; Kamber; Pei, 2011).

Já nos setores da indústria e da engenharia, a clusterização está diretamente associada ao controle de qualidade, processos e manutenção preventiva. Por meio da análise de dados de produção, o *clustering* permite o agrupamento de equipamentos detectando comportamentos semelhantes, além da detecção de padrões de falha de maneira antecipada (Tan; Steinbach; Kumar, 2014), contribuindo significativamente para a redução de custos e aprimoramento do sistema produtivo.

A compreensão das métricas de similaridade ou distância é essencial para os algoritmos de agrupamento (*clustering)*, pois permite quantificar o quão próximos ou distantes estão os elementos entre si. A escolha da métrica depende diretamente da natureza dos dados e dos objetivos esperados da análise. Para dados quantitativos, a distância euclidiana é mais comumente utilizada por sua simplicidade e interpretação geométrica direta (Jain; Murty; Flynn, 1999; Kaufman; Rousseeuw, 2005). Por outro lado, para dados categóricos, medidas alternativas, como a distância de Hamming, são mais adequadas por avaliarem a dissimilaridade com base nas diferenças entre atributos discretos (Hamming, 1950; Deza, M.; Deza, E., 2009). A definição apropriada da métrica influencia diretamente a formação dos *clusters* e a qualidade da segmentação resultante (Gao *et al*., 2023).

Os algoritmos de *clustering* apresentam características distintas, sendo adequados a diferentes tipos de conjuntos de dados e objetivos analíticos. De modo geral, esses algoritmos podem ser agrupados em categorias com base na forma como estruturam os agrupamentos. Entre os principais tipos, destacam-se: (i) algoritmos baseados em centróides, como o *K-Means*, que agrupam os dados em torno de pontos centrais representativos dos *clusters*; (ii) algoritmos baseados em densidade, como o DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*), que identificam agrupamentos a partir de regiões de alta densidade de dados, sendo capazes de lidar com ruídos e *clusters* de formatos arbitrários; (iii) algoritmos hierárquicos, como o *Agglomerative Clustering*, que constroem uma estrutura em forma de árvore (dendrograma), permitindo a análise em diferentes níveis de granularidade; e (iv) algoritmos baseados em distribuição, como os modelos de mistura gaussiana (*Gaussian Mixture Models*), que assumem que os dados são gerados a partir de distribuições probabilísticas subjacentes.

A escolha do algoritmo mais adequado depende das características dos dados analisados, como dimensionalidade, presença de ruído e formato esperado dos *clusters*, bem como dos objetivos da análise. No contexto desta pesquisa, optou-se pelo uso do algoritmo *K-Means* devido à sua eficiência computacional e à sua adequação à análise de dados textuais representados em espaços vetoriais de alta dimensão.

Os algoritmos de *clustering* podem ser organizados a partir de uma taxonomia baseada em suas estratégias de formação de grupos, destacando-se, entre as

principais categorias, os métodos hierárquicos, baseados em centróides, baseados em densidade e baseados em distribuição (Han; Kamber; Pei, 2011; Jain, 2010). Cada uma dessas abordagens apresenta características específicas quanto à forma de construção dos *clusters*, sensibilidade a ruídos e aplicabilidade a diferentes tipos de dados.

No contexto dos algoritmos hierárquicos, estes se caracterizam pela construção de uma estrutura em forma de árvore (dendrograma), que representa as relações de similaridade entre os dados em diferentes níveis de granularidade. Essa abordagem pode ser dividida em duas estratégias principais: aglomerativa (*bottom-up*) e divisiva (*top-down*).

Na abordagem aglomerativa, cada ponto de dados é inicialmente tratado como um *cluster* individual, sendo progressivamente agrupado a outros *clusters* com base em critérios de similaridade, até que todos os elementos estejam organizados em uma única estrutura hierárquica. Esse processo depende de métricas de distância e de critérios de ligação (*linkage*), como *single linkage*, *complete linkage* e *average linkage*, que influenciam diretamente a forma dos agrupamentos gerados.

Por outro lado, na abordagem divisiva, todos os dados são inicialmente considerados como pertencentes a um único *cluster*, sendo posteriormente particionados em subgrupos de forma sucessiva. Embora menos utilizada na prática devido ao seu maior custo computacional, essa estratégia pode oferecer melhor desempenho em cenários onde há separações mais evidentes entre os dados.

A principal diferença entre essas estratégias reside na forma de construção da estrutura de agrupamento: enquanto os métodos aglomerativos constroem os *clusters* por fusão progressiva de elementos, os métodos divisivos realizam sucessivas divisões de um conjunto inicial. Essa distinção impacta diretamente a complexidade computacional e a sensibilidade a ruídos e *outliers*.

Os métodos hierárquicos são particularmente úteis em análises exploratórias, pois permitem a visualização da estrutura dos dados em múltiplos níveis de granularidade, sendo amplamente aplicados em áreas como bioinformática, para análise de expressão gênica, e em segmentação de dados, quando há interesse na compreensão da estrutura interna dos agrupamentos (Everitt *et el.*, 2011).

Por outro lado, algoritmos não hierárquicos, como o *K-Means*, operam sob a premissa de um número predefinido de *clusters*. Este algoritmo aloca pontos de dados

de forma a minimizar a variância *intra-cluster* e maximizar a variância *inter-cluster*. Embora o *K-Means* seja amplamente utilizado devido à sua simplicidade e eficiência, ele pode ser limitado pela necessidade de especificar o número antecipadamente e pela sensibilidade à escolha dos centroides iniciais.

O algoritmo *K-Means* apresenta características específicas que influenciam diretamente sua aplicabilidade em diferentes contextos analíticos. Entre suas principais vantagens, destacam-se a simplicidade de implementação, a eficiência computacional, especialmente em grandes volumes de dados, e a facilidade de interpretação dos resultados, uma vez que os *clusters* são representados por centróides que sintetizam as características médias dos grupos.

Por outro lado, o método apresenta limitações relevantes. O *K-Means* exige a definição prévia do número de *clusters* (*k*), o que pode impactar significativamente os resultados quando não há conhecimento prévio da estrutura dos dados. Além disso, o algoritmo é sensível à inicialização dos centróides, podendo convergir para soluções locais subótimas, motivo pelo qual técnicas como o *K-Means++* são frequentemente utilizadas para melhorar a qualidade da inicialização.

Outra limitação importante refere-se à sua suposição implícita de *clusters* aproximadamente esféricos e de tamanho semelhante, o que pode reduzir seu desempenho em conjuntos de dados com estruturas mais complexas. Ademais, o algoritmo é sensível à presença de *outliers* e ruídos, que podem distorcer a posição dos centróides e comprometer a qualidade dos agrupamentos.

Apesar dessas limitações, o *K-Means* permanece amplamente utilizado em aplicações práticas, especialmente em cenários com dados de alta dimensionalidade e grande volume, como no caso da análise de dados textuais vetorizados, em que sua eficiência computacional e escalabilidade tornam-se fatores determinantes para sua adoção.

Assim, considerando que cada um desses métodos possui características, vantagens e limitações bastante perceptíveis, a escolha do algoritmo adequado se torna um aspecto crítico da análise (Mota *et al*., 2020).

Para determinar a eficácia de um agrupamento, métricas de validação interna podem ser utilizadas para quantificar, de forma simultânea, a coesão *intra-cluster* e a separação *inter-cluster*. Índices clássicos e amplamente empregados incluem a silhueta, que mede, para cada observação, o quanto ela está bem ajustada ao seu

próprio *cluster* em comparação com o *cluster* mais próximo (Rousseeuw, 1987), e o índice Davies–Bouldin, que combina medidas de dispersão *intra-cluster* com distâncias entre centroides para avaliar a separabilidade dos grupos (Davies; Bouldin, 1979). Em geral, recomenda-se avaliar a qualidade do agrupamento usando múltiplos índices e complementar essa avaliação com análise visual e testes estatísticos, pois cada índice enfatiza aspectos distintos da estrutura dos dados e pode ter sensibilidades diferentes a formatos (Kaufman; Rousseeuw, 2005; Arbelaitz *et al*., 2013).

Uma categoria distinta é representada pelos algoritmos baseados em densidade, como o *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN). Esses métodos identificam regiões de alta densidade de pontos de dados, separados por regiões de baixa densidade. Sua capacidade de identificar *clusters* de formas arbitrárias e lidar com *outliers* torna-os particularmente úteis em aplicações complexas (Ventorim; Luchi; Varejão, 2020).

Os algoritmos espectrais oferecem uma abordagem baseada na teoria dos grafos, onde a estrutura dos dados é capturada através de um gráfico e o *clustering* é realizado através da análise das propriedades espectrais desse gráfico. Conforme Nascimento e Carvalho (2011), este método é eficaz especialmente em dados que não são isotrópicos e as relações entre os pontos de dados são melhor representadas em um espaço de menor dimensão.

No contexto de um processo decisório, como se apresenta na presente pesquisa, a avaliação da eficiência é um componente crítico. Essa avaliação tem como finalidade assegurar que os agrupamentos gerados pelos algoritmos são, de fato, significativos e úteis para a análise de dados. A eficiência pode ser entendida como a capacidade de um algoritmo de formar *clusters* que são internamente coesos e externamente distintos uns dos outros.

A compacidade refere-se ao grau em que os membros estão próximos uns dos outros, indicando uma alta similaridade interna. A separação, por outro lado, diz respeito à distância entre diferentes *clusters*, refletindo a dissimilaridade entre grupos. Uma boa partição deve maximizar tanto a compacidade interna quanto a separação entre os grupos.

Métricas específicas foram desenvolvidas para quantificar esses conceitos. Por exemplo, o coeficiente de silhueta mede tanto a coesão quanto a separação de um

cluster, fornecendo uma avaliação balanceada. Segundo Tan, Steinbach e Kumar (2014), além do método do cotovelo, outras métricas são amplamente utilizadas para avaliar a qualidade dos agrupamentos, destacando-se o índice Davies-Bouldin e o índice Dunn, os quais incorporam diferentes critérios de coesão *intra-cluster* e separação *inter-cluster*.

O índice Davies-Bouldin baseia-se na razão entre a dispersão interna dos *clusters* e a distância entre seus centróides, sendo calculado a partir da média das similaridades entre cada *cluster* e aquele que lhe é mais semelhante. Valores mais baixos desse índice indicam melhor qualidade de agrupamento, pois refletem *clusters* mais compactos e bem separados entre si.

Por sua vez, o índice Dunn é definido como a razão entre a menor distância *inter-cluster* e a maior distância *intra-cluster*, buscando maximizar a separação entre os grupos ao mesmo tempo em que minimiza sua dispersão interna. Diferentemente do índice Davies-Bouldin, valores mais elevados do índice Dunn indicam melhor qualidade dos *clusters*.

A utilização conjunta dessas métricas permite uma avaliação mais robusta dos resultados, uma vez que cada uma enfatiza diferentes aspectos da estrutura dos dados, contribuindo para a validação dos agrupamentos obtidos.

Realizado o agrupamento dos *clusters*, é importante avaliar se as diferenças percebidas entre eles são relevantes, estatisticamente falando. Uma ferramenta possível para validação é a análise de variância (ANOVA), que pode ser empregada para verificar se as médias das variáveis analisadas diferem entre os grupos formados, garantindo que as distinções encontradas não sejam apenas produtos do acaso, mas reflitam uma estrutura inerente aos dados (Hair *et al*., 2019; Field, 2013). Essa etapa é essencial para validar empiricamente os resultados do agrupamento, contribuindo para a robustez e credibilidade das interpretações obtidas (Coombes *et al*., 2021).

Os tipos de algoritmos ainda se distinguem pelas características das estruturas de dados. Algoritmos baseados em densidade, como DBSCAN, por exemplo, são eficazes em detectar *clusters* de formas arbitrárias, mas podem não ser ideais em informações com variações de densidade. Estudos realizados por Hahsler, Piekenbrock e Doran (2019) mostram que a seleção inadequada do algoritmo pode

levar a agrupamentos que não representam a estrutura real dos dados, comprometendo a eficácia da análise.

Os algoritmos podem ser categorizados ainda pelo tipo de treinamento, podendo ser supervisionado e não supervisionado. No treinamento supervisionado, os modelos são treinados com dados rotulados, permitindo a predição de classes ou valores para novos dados. Essa abordagem é amplamente utilizada em tarefas de classificação e regressão. Em contrapartida, o treinamento não supervisionado não depende de rótulos, buscando identificar estruturas subjacentes ou agrupamentos nos dados, como ocorre no *clustering*.

Estudos de Jain (2010) e Aggarwal (2018) evidenciam as vantagens e limitações de cada abordagem. Enquanto o treinamento supervisionado tende a oferecer maior precisão quando há disponibilidade de dados rotulados, o treinamento não supervisionado permite a descoberta de padrões ocultos e a exploração de dados sem uma classificação prévia.

Alguns algoritmos podem ser excessivamente influenciados por *outliers*, resultando em agrupamentos que não refletem adequadamente a distribuição geral dos dados. Segundo Everitt *et al*. (2011), abordagens robustas ou métodos prévios de tratamento de *outliers* podem ser necessários para superar esse problema.

O *clustering*, enquanto técnica de análise exploratória de dados, apresenta ampla aplicabilidade em contextos organizacionais, especialmente na identificação de padrões e segmentação de informações relevantes para a tomada de decisão. Ao agrupar elementos com características semelhantes, essa abordagem permite transformar grandes volumes de dados em estruturas interpretáveis, facilitando a identificação de comportamentos recorrentes e perfis distintos.

No contexto organizacional, essa capacidade pode ser aplicada, por exemplo, na segmentação de clientes, na identificação de padrões de consumo, na detecção de anomalias em processos operacionais e na análise de documentos administrativos. Dessa forma, o *clustering* contribui para a geração de conhecimento a partir dos dados, apoiando decisões estratégicas e operacionais baseadas em evidências.

Em contextos organizacionais, o *clustering* é amplamente utilizado como ferramenta de segmentação e identificação de padrões, permitindo apoiar decisões estratégicas em diferentes áreas funcionais. Uma de suas aplicações mais consolidadas ocorre no campo do marketing, em que a técnica é empregada para

categorizar clientes em grupos com características semelhantes, possibilitando a personalização de estratégias, o aumento da eficácia de campanhas e a melhoria da satisfação do cliente (Melo *et al.*, 2022).

De forma complementar, o *clustering* também é aplicado na análise de padrões de consumo, permitindo identificar produtos frequentemente adquiridos em conjunto. Essas informações subsidiam estratégias de *cross-selling* e *up-selling*, além de contribuírem para o gerenciamento eficiente de estoques e para a otimização da disposição de produtos em ambientes físicos e digitais, aprimorando a experiência do consumidor (Aguiar; Sampaio, 2014).

No âmbito da gestão de recursos humanos, a técnica possibilita a identificação de padrões relacionados ao comportamento organizacional, incluindo desempenho, satisfação e rotatividade de colaboradores. A partir dessas análises, torna-se possível desenvolver políticas mais eficazes de retenção de talentos e aumento da produtividade (Sakka *et al.*, 2022).

Adicionalmente, na área de operações, o *clustering* contribui para a otimização da cadeia de suprimentos, ao permitir o agrupamento de fornecedores e clientes com base em variáveis como localização, volume e frequência de pedidos. Essa segmentação possibilita a melhoria do planejamento logístico, a redução de custos de transporte e o aumento da eficiência operacional (Shi, 2024).

2.2.2 Algoritmo *K-Means*

O algoritmo *K-Means* é um dos métodos de *clustering* mais amplamente utilizados na literatura de aprendizado de máquina (*machine learning*), sendo classificado como um algoritmo de particionamento baseado em centróides. Seu objetivo principal é dividir um conjunto de dados em *k* grupos distintos, de modo que os elementos pertencentes a um mesmo *cluster* apresentem alta similaridade entre si, enquanto elementos de *clusters* diferentes sejam o mais distintos possível (Jain, 2010).

Do ponto de vista formal, o *K-Means* busca minimizar a variabilidade interna dos *clusters*, frequentemente representada pela soma dos quadrados das distâncias entre os pontos e o centróide de seu respectivo grupo (*Within-Cluster Sum of Squares* – *WCSS*). Esse processo pode ser interpretado como um problema de otimização, no

qual se busca encontrar a melhor configuração de agrupamento para um número previamente definido de *clusters*.

O *K-Means* possui uma longa e diversa história, tendo sido inicialmente concebido em campos científicos distintos. Conforme destacado por Jain (2010), a origem do algoritmo está vinculada a Steinhaus (1956), que propôs a ideia do particionamento de conjuntos de dados tendo por base a proximidade entre os pontos; em seguida Lloyd (1957), visando o uso em técnicas de compressão de sinais, propôs o algoritmo como é conhecido. O termo *"K-Means"* somente foi estabelecido em 1967 por MacQueen (1967), sendo também conhecido como "Algoritmo de Lloyd" ou "Algoritmo de Lloyd-Forgy", tendo em vista que a ideia foi publicada em 1965 por Edward W. Forgy (Forgy, 1965).

A literatura atual de *machine learning* estabelece o algoritmo *K-Means* como um dos métodos mais difundidos para a realização de análises de agrupamentos. Este método se caracteriza por possuir um alto grau de simplicidade conceitual, ser eficiente computacionalmente e robusto em aplicações de larga escala (Jain; Murty; Flynn, 1999). Se trata de um método de aprendizado não supervisionado, que possui como objetivo o particionamento de um conjunto de dados em *k* grupos mutuamente exclusivos, visando minimizar a variabilidade interna do grupo e potencializar e heterogeneidade entre estes (Kaufman; Rousseeuw, 2005). O objetivo principal do *K-Means* está em minimizar a soma das distâncias quadráticas entre os pontos e o centroide do *cluster* em que pertencem. Esta métrica é conhecida como *Sum of Squared Errors* (SSE).

O *K-Means* opera de forma iterativa. De início, são selecionados *k* centroides (de maneira aleatória ou baseando-se em uma regra). Os elementos são então atribuídos ao centroide mais próximo, baseando-se em uma medida de distância, a exemplo da distância euclidiana. Em seguida, os centroides são recalculados tendo por base a média dos pontos pertencentes a cada grupo. Esse mesmo processo é então repetido até que a convergência seja alcançada, ou seja, quando não houver mais alterações significativas na atribuição dos dados ou nos centroides (Gao *et al*., 2023).

O algoritmo *K-Means* é utilizado de ampla em diversas áreas, o que demonstra a versatilidade de uso e robustez da ferramenta. Na gestão de operações, o algoritmo pode ser utilizado para a segmentação de fornecedores, resultando em uma

otimização da cadeia de suprimentos, possibilitando identificar agrupamentos de fornecedores, orientando estratégias de melhoria (Shi, 2024). Já Mehta *et al*. (2025) demonstra em seus estudos que o algoritmo *K-Means* pode ser aplicado na área de finanças utilizando-se aprendizado não supervisionado, visando possibilitar a identificação de padrões de fraude, permitindo a manutenção da integridade e estabilidade do sistema financeiro.

De maneira geral, o *K-Means* possui vantagens notáveis, como eficiência computacional, escalabilidade e facilidade de interpretação dos dados. Já as limitações mais perceptíveis são: a sensibilidade à escolha dos centroides, desempenho reduzido com *clusters* não esféricos ou na presença de ruído, a necessidade de definição prévia do número de *clusters*.

O algoritmo *K-Means* é um dos métodos de *clustering* mais conhecidos e amplamente utilizados devido à sua simplicidade conceitual e eficiência computacional (Kanungo *et al*., 2002). Seu objetivo matemático é particionar um conjunto de n observações em *k clusters*, de modo a minimizar a variância intra-*cluster*, que é formalmente definida como a Soma das Distâncias Quadráticas (SSE - *Sum of Squares Errors*). A função objetivo a ser minimizada é expressa pela seguinte equação:

$$argmin_{\mathrm{s}} \sum_{\mathrm{i}=1}^{\mathrm{k}} \sum_{\mathrm{x}} \in S_{\mathrm{i}} \|x - \mu_{\mathrm{i}}\|^2$$

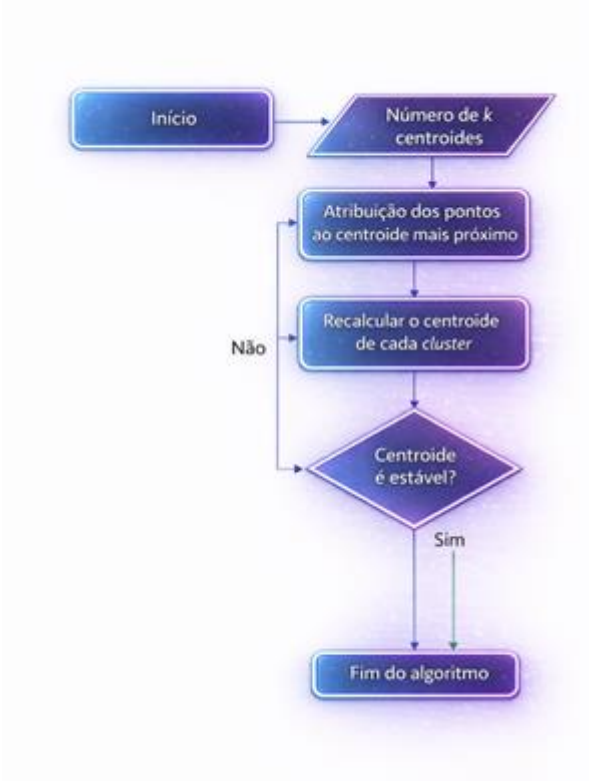
Nesta equação, $S_i$representa o conjunto de pontos de dados pertencentes ao *cluster* i, e μ_i é o centroide (a média vetorial) de todos os pontos emS_i. Essencialmente, o algoritmo busca encontrar os centroides e as atribuições de *cluster* que tornam os grupos o mais compactos possível, medido pela distância euclidiana quadrada.

Para resolver este problema de otimização, o *K-Means* utiliza um algoritmo iterativo heurístico conhecido como algoritmo de *Lloyd*. Este processo alterna entre duas etapas principais até que a convergência seja alcançada:

- Etapa de Atribuição (*Assignment Step*): Nesta fase, cada ponto de dados é atribuído ao *cluster* cujo centroide está mais próximo. A proximidade é calculada usando a distância euclidiana.

- Etapa de Atualização (*Update Step*): Após todos os pontos terem sido atribuídos a um *cluster*, o centroide de cada *cluster* é recalculado como a média de todos os pontos de dados que foram atribuídos a ele.

Figura 4 - Fluxograma de Execução do Algoritmo *K-Means*



Fonte: elaborado pelo autor (2026)

Essas duas etapas são repetidas iterativamente. A cada iteração, o valor da função objetivo SEE diminui, garantindo que o algoritmo eventualmente convirja. A convergência ocorre quando as atribuições de *cluster* não mudam mais entre as iterações, ou quando um número máximo de iterações é atingido. Apesar de sua simplicidade e velocidade, o *K-Means* padrão possui desvantagens notáveis: a necessidade de pré-especificar o número de *clusters k*, sua sensibilidade à escolha inicial dos centroides, o que pode levá-lo a convergir para mínimos locais em vez do ótimo global, e sua dificuldade em lidar com clusters de formas não esféricas, densidades variadas ou a presença de *outliers*.

2.2.2.1 Aprimorando a inicialização com *K-Means++*

Uma das fraquezas mais significativas do algoritmo *K-Means* padrão é sua sensibilidade à seleção inicial dos centroides. Uma inicialização aleatória inadequada, onde os centroides iniciais são escolhidos muito próximos uns dos outros ou em regiões esparsas do espaço de dados, pode fazer com que o algoritmo de *Lloyd* convirja para um mínimo local de baixa qualidade, resultando em um agrupamento que não reflete a estrutura natural dos dados. Para mitigar este problema, Arthur e Vassilvitskii (2007) propuseram um método de inicialização aprimorado chamado *K-Means++*.

O *K-Means++* substitui a inicialização puramente aleatória por um procedimento de "semeadura cuidadosa" (*careful seeding*) que distribui os centroides iniciais de forma mais inteligente pelo conjunto de dados. A intuição por trás do método é que os centroides iniciais devem estar bem espaçados entre si. O algoritmo funciona da seguinte maneira:

- O primeiro centroide, c1, é escolhido uniformemente ao acaso a partir do conjunto de pontos de dados X.
- Para cada ponto de dados $x \in X$, calcula-se D(x), a distância euclidiana entre x e o centroide mais próximo que já foi escolhido.
- O próximo centroide, ci, é escolhido a partir dos pontos de dados restantes com uma probabilidade proporcional a D(x)2. Ou seja, um ponto é escolhido com uma probabilidade $\sum x' \in X D(x')2 D(x)2$. Este passo de ponderação probabilística favorece a seleção de pontos que estão longe dos centroides já existentes.
- Os passos 2 e 3 são repetidos até que todos os *k* centroides tenham sido escolhidos.

A principal contribuição teórica do *K-Means++* é que ele oferece uma garantia de aproximação. Enquanto o *K-Means* padrão pode produzir um agrupamento com um erro arbitrariamente grande em comparação com a solução ótima, o *K-Means++* garante que a solução encontrada é, em expectativa, *O(logk)-competitiva* com a solução ótima de *K-Means*. Na prática, essa inicialização mais inteligente não apenas

melhora a qualidade do agrupamento final, mas também acelera a convergência do algoritmo de *Lloyd*, muitas vezes reduzindo o número total de iterações necessárias.

2.2.2.2 Método do Cotovelo (*Elbow Method)*

A definição do número adequado de *clusters* (*k*) é uma etapa crítica na aplicação do algoritmo *K-Means*, uma vez que influencia diretamente a qualidade e a interpretabilidade dos agrupamentos obtidos. Nesse contexto, diferentes métodos podem ser utilizados para auxiliar na determinação do valor mais apropriado de *k*, entre os quais se destacam o método da silhueta e o Método do Cotovelo, do inglês "*Elbow Method"*, proposto por Thorndike (1953).

O método da silhueta (*Silhouette Method*) é uma técnica amplamente utilizada para avaliar a qualidade dos agrupamentos, considerando simultaneamente a coesão interna dos *clusters* e a separação entre eles. Para cada ponto de dados, calcula-se o coeficiente de silhueta, que varia no intervalo de -1 a 1, sendo definido a partir da comparação entre a distância média do ponto aos demais elementos do mesmo *cluster* e a distância média aos pontos do *cluster* mais próximo.

Valores próximos de 1 indicam que o ponto está bem alocado em seu *cluster*, apresentando alta similaridade com os elementos do mesmo grupo e baixa similaridade com outros *clusters*. Valores próximos de 0 indicam que o ponto se encontra em uma região de transição entre *clusters*, enquanto valores negativos sugerem possível alocação incorreta.

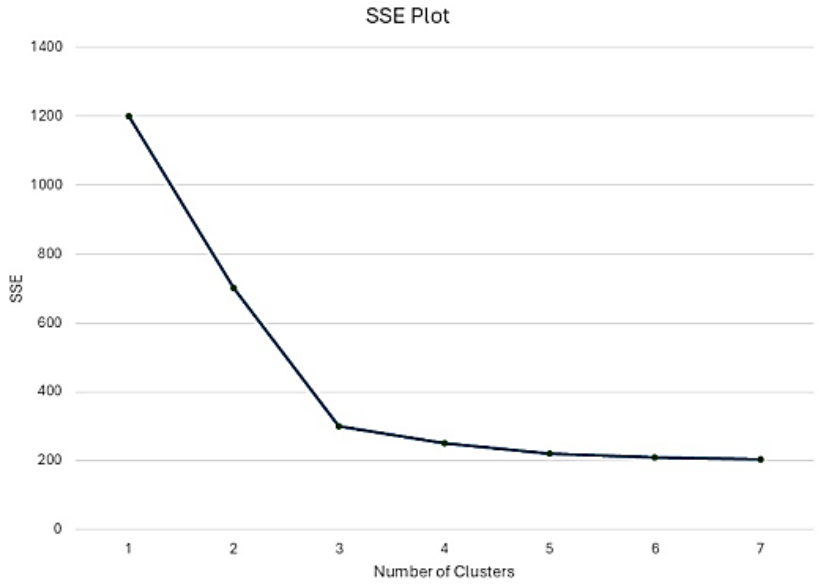
A média dos coeficientes de silhueta para todos os pontos fornece uma medida global da qualidade do agrupamento, sendo utilizada como critério para seleção do número de *clusters*. Nesse sentido, o valor de *k* que maximiza o coeficiente médio de silhueta tende a representar uma configuração mais adequada dos dados.

Apesar de sua robustez, o método da silhueta pode apresentar limitações em conjuntos de dados de alta dimensionalidade, como aqueles derivados de representações vetoriais de textos, nos quais as distâncias podem se tornar menos discriminativas. Nessas situações, sua interpretação pode ser comprometida, sendo recomendável sua utilização em conjunto com outros métodos de avaliação.

Complementarmente, o método do cotovelo (*Elbow Method*) é amplamente utilizado como uma abordagem prática para definição do número de *clusters*,

especialmente em cenários de grande volume de dados. Esse método baseia-se na análise da redução da variabilidade interna dos *clusters* à medida que o número de grupos aumenta. O procedimento consiste na execução sucessiva do *K-Means*, variando o número de *clusters k*, registrando-se os valores da soma das distâncias quadráticas (SSE) em cada execução. À medida em que ocorre o aumento de *k*, a SSE tenderá a decrescer de forma rápida até determinado ponto, a partir do qual se tornará marginal. Esse ponto de inflexão, que graficamente se assemelha a um cotovelo, indica o número ótimo de *clusters*, visto que representa a ocorrência de equilíbrio entre a complexidade do modelo e a explicação da variabilidade dos dados. O estudo apresentado por Herdiana *et al*. (2025) demonstra visualmente a ocorrência de SSE com curva clara.

Figura 5 - Visualização Gráfica de SSE com Curva Clara



Fonte: Herdiana *et al*. (2025)

Na presente pesquisa, o método do cotovelo foi escolhido em função de permitir uma seleção objetiva de *clusters k*, minimizando que ocorra o enviesamento subjetivo, além de contribuir para uma maior consistência analítica na identificação de agrupamentos dos padrões existentes nos dados textuais analisados, oriundos dos documentos licitatórios.

2.2.3 Classificação Textual

A classificação textual se apresenta como uma das aplicações mais significativas do processamento de linguagem natural (PLN), baseando-se em uma metodologia de atribuição automatizada de categorias ou rótulos em documentos conforme padrões linguísticos identificados no conteúdo dos mesmos. Segundo Jurafsky e Martin (2014), a classificação textual possibilita transformar dados textuais sem estruturação em representações computacionais tratáveis, o que permite a construção de modelos preditivos com capacidade de aprendizado de relações semânticas e identificação de tendências que seriam impossíveis de se identificar de forma manual. Em complemento, Allahyari *et al*. (2017) destacam que a classificação textual faz parte do núcleo de tarefas do tipo *text mining*, se mostrando como uma ferramenta fundamental para a filtragem automatizada de documentos, gestão da informação e suporte no processo decisório.

Ao permitir uma análise automatizada de volumes gigantescos de dados e documentos, a classificação textual reduz de forma significativa a complexidade de informações e possibilita que sejam detectados padrões de maneira escalável e eficiente. Allahyari *et al*. (2017) e Jurafsky e Martin (2014) demonstram que algoritmos de aprendizado supervisionado aplicados na classificação textual, como Máquina de Vetores de Suporte (do inglês, SVM) e redes neurais, possuem a capacidade de criar classificações com elevado poder preditivo, em especial nos domínios caracterizados por grande variação textual.

Dentre as fases da implementação de técnicas de classificação e agrupamento de documentos, está o pré-processamento textual. Esta etapa é essencial e visa transformar os dados brutos em uma estrutura tratável de forma computacional. Inicialmente é realizada uma limpeza dos dados, com a remoção de caracteres especiais, espaços em excesso, números e outros elementos que não irão contribuir para a extração do significado semântico. Nesta etapa também pode ocorrer a correção automática de grafia e normalização de termos, o que se mostra especialmente relevante para a análise de documentos jurídicos, como os que são objeto da presente pesquisa.

Em seguida, realiza-se a *tokenização*, onde o texto é segmentado em unidades menores, as quais servirão de base para análise. Na etapa seguinte ocorre a remoção

de *stopwords*, como artigos, preposições e pronomes (“de”, “para”, “pela”, “deste”, “na”, entre outras), palavras que ocorrem frequentemente, mas com baixo conteúdo. Por fim, na lematização (*lemmatization*), ocorre a redução das palavras às suas formas canônicas, preservando-se o sentido linguístico e eliminando variações de flexão. Essa etapa contribui para a obtenção de um vocabulário mais uniforme e para a redução da dimensionalidade dos dados.

Para a conversão dos textos em representações numéricas, há uma grande gama de métodos disponíveis, dentre os quais se destacam os de representação esparsa, como o *Bag of Words (BoW)*, que se baseia na frequência dos termos. Outra opção bastante utilizada é o *Term Frequency-Inverse Document Frequency (TF-IDF)*, que também considera a frequência das palavras, mas pondera sua relevância no contexto dos dados.

Outros métodos mais avançados vêm sendo utilizados como alternativa para a detecção de relações semânticas entre palavras. Nesse contexto, os *word embeddings* representam uma evolução significativa em relação às abordagens tradicionais baseadas em representações esparsas, como o modelo *bag-of-words*, ao produzirem vetores densos e contínuos em espaços de menor dimensionalidade (Mikolov *et al.*, 2013; Pennington; Socher; Manning, 2014).

Modelos como o *Word2Vec* (Mikolov *et al*., 2013) e o GloVe (Pennington; Socher; Manning, 2014) permitem capturar relações semânticas e sintáticas entre palavras, de modo que termos utilizados em contextos semelhantes apresentem representações vetoriais próximas no espaço semântico. Essa propriedade contribui para o aprimoramento do desempenho em diversas tarefas de processamento de linguagem natural, incluindo classificação textual e agrupamento de documentos.

No contexto desta pesquisa, a utilização de *word embeddings* possibilita representar documentos licitatórios de forma mais informativa e semanticamente consistente, favorecendo a identificação de padrões relevantes nos dados e contribuindo para a qualidade dos agrupamentos obtidos.

Essas abordagens permitem superar limitações das representações esparsas, especialmente no que se refere à captura de similaridades semânticas em grandes corpora textuais.

Quadro 1 - Análise Comparativa de Modelos de *Word Embedding*

Característica	***Word2Vec***	**GloVe (*Global Vectors*)**	***FastText***
Modelo Subjacente	Preditivo (Rede Neural)	Baseado em Contagem (Fatorização de Matriz)	Preditivo (baseado em subpalavras)
Contexto Utilizado	Local (janela deslizante de palavras)	Global (matriz de coocorrência de todo o corpus)	Local (com informação de subpalavras)
Morfologia	Trata palavras como unidades atômicas	Trata palavras como unidades atômicas	Lida bem com morfologia rica (n-gramas de caracteres)
Palavras OOV	Não consegue gerar vetores para palavras fora do vocabulário	Não consegue gerar vetores para palavras fora do vocabulário	Consegue gerar vetores para palavras fora do vocabulário
Vantagem Principal	Eficiente, bom em capturar analogias semânticas	Bom desempenho com corpus menores, captura estatísticas globais	Robusto a erros de digitação e palavras raras/novas

Fonte: elaborada pelo autor (2026)

Quadro 2 - Comparação entre Representações Esparsas e Densas

Característica	**Representações Esparsas (BoW, TF-IDF)**	**Representações Densas (*Word2Vec*, *GloVe*)**
Forma de Representação	Baseada na frequência dos termos	Detecta relações semânticas
Dimensionalidade	Alta, proporcional ao tamanho do vocabulário	Reduzida, girando em torno de 50 a 300 dimensões
Correlação entre Palavras	Não capta similaridades entre termos	Palavras com semântica próxima possuem vetores semelhantes
Sensibilidade a Ruídos	Alta sensibilidade a ruídos	Sensibilidade menor a ruídos
Exemplos de Aplicação	Sistemas de recuperação de informações, análise de frequência	Análise de similaridade textual, agrupamento temático, modelos preditivos
Aplicabilidade para *clusters*	Menos eficaz em tarefas semânticas	Mais eficaz no agrupamento pela semântica

Fonte: elaborada pelo autor (2026)

Considerando o contexto da presente pesquisa, a aplicação de técnicas de classificação textual nos processos de licitação possibilita a identificação automatizada de documentos com maior propensão a ocorrência de desequilíbrios econômico-financeiros, de forma a agrupar os mesmos em *clusters* conforme semelhanças entre os mesmos, contribuindo significativamente para a implementação de mecanismos de predição que visem fortalecer a eficiência administrativa e a governança pública, auxiliando diretamente no processo decisório.

2.2.3.1 Inteligência de dados textuais em aquisições

A transformação de vastos repositórios de documentos de licitação, inerentemente não estruturados e redigidos em linguagem natural, em um formato quantitativo passível de análise por algoritmos de aprendizado de máquina, exige a construção de um arcabouço metodológico rigoroso e multifásico. Este processo, conhecido como *pipeline* de Processamento de Linguagem Natural (PLN), compreende uma sequência de etapas interdependentes, cada uma com seu próprio fundamento teórico e desafios práticos. A escolha das técnicas em cada fase, desde a extração do texto bruto até sua representação final como vetores numéricos, assim das técnicas a serem utilizadas, impacta diretamente a qualidade dos padrões que podem ser descobertos e, consequentemente, a validade das conclusões estratégicas (Lakatos *et al*., 2024). Esta seção detalha a arquitetura técnica do *pipeline* desenvolvido para esta pesquisa, justificando cada decisão metodológica à luz dos princípios da ciência de dados e das especificidades do domínio da contratação.

2.2.3.1.1 O uso do processamento de linguagem natural para análise documental

O Processamento de Linguagem Natural (PLN) é um ramo da inteligência artificial que se concentra em capacitar as máquinas a entender, interpretar e manipular a linguagem humana (Nadkarni *et al*., 2011). No contexto da administração pública, onde a produção documental é massiva e o conteúdo é denso em terminologia técnica e jurídica, o PLN oferece um conjunto de ferramentas indispensáveis para automatizar a análise e extrair *insights* que seriam inviáveis por meio de leitura manual. A capacidade de um sistema de PLN de lidar com dialetos,

jargões e irregularidades gramaticais o torna particularmente adequado para decifrar a complexidade dos textos de licitação.

O *pipeline* de PLN, ou seja, a sequência de operações para processar dados textuais, geralmente segue uma estrutura padronizada (Naseem *et al*., 2020). Inicia-se com a aquisição e extração dos dados brutos de suas fontes originais. Segue-se uma fase crítica de pré-processamento e normalização, na qual o texto é limpo e padronizado para remover ruídos e inconsistências. A etapa subsequente, conhecida como engenharia de características ou vetorização, é responsável por transformar o texto limpo em uma representação numérica. É somente após essa transformação que os algoritmos de aprendizado de máquina, como os de *clustering* ou de classificação, podem ser aplicados. Finalmente, os resultados do modelo são avaliados e interpretados para gerar conhecimento acionável.

3 METODOLOGIA

3.1 FUNDAMENTAÇÃO DA ABORDAGEM METODOLÓGICA

A definição da abordagem metodológica representa uma etapa fundamental na estruturação de qualquer pesquisa científica, pois orienta os procedimentos adotados para coleta, tratamento e análise dos dados, ao mesmo tempo em que estabelece os limites epistemológicos da investigação. A distinção entre o conhecimento científico e o senso comum reside essencialmente na existência de critérios sistemáticos de controle, verificação e replicabilidade. Como apontam Sampieri, Collado e Lucio (2013), a pesquisa científica é composta por um conjunto de processos sistemáticos e empíricos utilizados com o propósito de compreender e explicar fenômenos de forma rigorosa.

Complementando essa perspectiva, Alves (2004) argumenta que a ciência não é uma nova forma de conhecimento, mas sim uma intensificação das capacidades cognitivas comuns a todos os indivíduos, refinadas por meio de instrumentos de controle e rigor que viabilizam sua especialização. Assim, o diferencial do conhecimento científico repousa na sua estrutura lógica e sistemática, que assegura precisão, consistência e confiabilidade às inferências realizadas.

A escolha da abordagem metodológica mais adequada está diretamente relacionada à natureza do problema de pesquisa e aos objetivos do estudo. Fatores como a complexidade do fenômeno investigado, a natureza das variáveis e a possibilidade de mensuração objetiva são determinantes nesse processo decisório (Creswell, 2014; Gil, 2008).

Nesse contexto, é necessário considerar as principais abordagens metodológicas adotadas na pesquisa científica: qualitativa, quantitativa e mista. Cada uma dessas abordagens apresenta características específicas quanto à forma de coleta, análise e interpretação dos dados, sendo selecionadas de acordo com as demandas e características do objeto de estudo (Lakatos; Marconi, 2017; Creswell, 2014).

A definição da abordagem metodológica constitui etapa fundamental do delineamento da pesquisa, influenciando diretamente os procedimentos de coleta e análise dos dados.

Cada uma dessas abordagens parte de pressupostos distintos. A abordagem qualitativa busca compreender fenômenos em sua totalidade, enfatizando contextos subjetivos e significados atribuídos pelos indivíduos. A abordagem quantitativa, por sua vez, pauta-se pela objetividade, pela mensuração de variáveis e pela análise estatística dos dados. Já a abordagem mista integra os pressupostos das abordagens anteriores, proporcionando uma análise combinada de dados numéricos e descritivos.

Embora partam de fundamentos epistemológicos distintos, ambas as abordagens qualitativa e quantitativa compartilham, conforme Grinnell (2005), um conjunto de etapas fundamentais no processo de construção do conhecimento científico:

- Observação e avaliação de fenômenos;
- Formulação de suposições ou hipóteses;
- Teste das hipóteses com base em dados;
- Revisão das hipóteses à luz das evidências empíricas;
- Proposição de novas observações e inferências, visando ao aprimoramento do conhecimento.

A presente pesquisa adota a abordagem quantitativa por ser mais compatível com os objetivos delineados e com a natureza do fenômeno a ser investigado. Segundo Sampieri, Collado e Lucio (2013), a abordagem quantitativa fundamenta-se na coleta sistemática de dados numéricos, com o propósito de testar hipóteses e estabelecer relações entre variáveis por meio de procedimentos estatísticos. Esta abordagem permite comprovar teorias, identificar padrões e avaliar a consistência das relações causais observadas.

Bryman (2015) destaca que os principais fundamentos da abordagem quantitativa se concentram na mensuração, na verificação de causalidade, na generalização dos resultados e na replicação das análises. Tais fundamentos são detalhados por Martins (2012), que apresenta as seguintes características essenciais dessa abordagem:

- Mensurabilidade: exige variáveis bem definidas e passíveis de quantificação;
- Causalidade: permite identificar relações de causa e efeito entre variáveis;

- Generalização: possibilita a extrapolação dos resultados obtidos para além da amostra estudada;
- Replicação: assegura a reprodutibilidade da pesquisa por outros pesquisadores, conferindo validade externa aos resultados.

Ainda conforme Sampieri, Collado e Lucio (2013), o processo investigativo na abordagem quantitativa é sequencial, dedutivo, lógico e comprobatório, voltado à análise objetiva da realidade. Seus benefícios incluem: controle sobre os fenômenos estudados, precisão na mensuração, possibilidade de previsão de comportamentos e replicação de resultados.

Dadas essas características, e considerando que a presente pesquisa objetiva analisar de forma objetiva os dados relacionados às contratações da Universidade Federal de Mato Grosso (UFMT), buscando identificar padrões, categorizar dados e estabelecer relações causais entre as variáveis envolvidas, a abordagem quantitativa mostra-se como a mais apropriada para atender ao escopo da investigação.

3.2 JUSTIFICATIVA DO MÉTODO: MODELAGEM E SIMULAÇÃO

Tendo sido definida a abordagem quantitativa como diretriz epistemológica da pesquisa, o passo subsequente consiste na seleção do método mais adequado à operacionalização do estudo. Seguindo a classificação proposta por Martins (2012), os principais métodos aplicáveis a pesquisas de natureza quantitativa são: pesquisa de avaliação (*survey*); modelagem e simulação; experimento; e quase-experimento.

Dentre essas possibilidades, a presente pesquisa adota como método a modelagem e simulação, por apresentar elevada aderência aos objetivos do estudo e à natureza do problema investigado. A análise dos processos licitatórios envolve múltiplos fatores de natureza técnica, operacional e econômica, expressos, em grande parte, em documentos textuais não estruturados. Essa característica dificulta a identificação de padrões relevantes por meio de abordagens tradicionais, baseadas em análises manuais ou regras previamente definidas.

Nesse contexto, torna-se necessária a utilização de técnicas capazes de extrair e organizar informações a partir de grandes volumes de dados textuais, permitindo identificar similaridades entre processos e padrões associados à ocorrência de

desequilíbrio econômico-financeiro. Dessa forma, a aplicação de métodos de processamento de linguagem natural e algoritmos de agrupamento mostra-se adequada para apoiar a análise proposta nesta pesquisa, contribuindo para a geração de informações estruturadas a partir de dados não estruturados.

A modelagem quantitativa, nesse cenário, constitui um instrumento eficaz para representar os processos reais, testar diferentes cenários e antever os impactos decorrentes de cada alternativa decisória. Em linhas de produção, por exemplo, a modelagem permite responder de forma estruturada a questões como: o que produzir, quanto produzir, quando produzir e como produzir. Tais decisões impactam diretamente variáveis críticas, como tempo de ciclo, custos operacionais e nível de atendimento da demanda. Além disso, aplicações em áreas como controle de estoques, logística e planejamento de capacidade também se beneficiam amplamente da modelagem e simulação.

Conforme Bertrand e Fransoo (2002), o uso da modelagem quantitativa em pesquisa operacional consolidou-se a partir da década de 1960, inicialmente com foco na resolução de problemas idealizados, situações simplificadas da realidade, onde apenas os elementos considerados essenciais eram incorporados ao modelo. Essa abordagem deu origem ao que Morabito Neto e Pureza (2012) denominam pesquisa axiomática, caracterizada por proposições analíticas formuladas com base em abstrações controladas.

Com o avanço das técnicas estatísticas, matemáticas e computacionais, houve uma ampliação progressiva da complexidade dos modelos, aproximando-os dos problemas reais enfrentados nas organizações. Essa transição marca o surgimento da pesquisa empírica, definida por Morabito Neto e Pureza (2012) como aquela que se vale de dados concretos para formular, validar e aplicar modelos quantitativos voltados à resolução de problemas do mundo real.

Em situações nas quais a complexidade do sistema inviabiliza uma modelagem matemática analítica, a simulação computacional surge como alternativa metodológica viável. De acordo com Bertrand e Fransoo (2002), a simulação permite representar, por meio de algoritmos e poder computacional, o comportamento de sistemas complexos ao longo do tempo, considerando um número elevado de variáveis de entrada. Morabito Neto e Pureza (2012) destacam ainda que a simulação

é especialmente útil para analisar a evolução de processos sob diferentes condições operacionais, permitindo testar hipóteses, validar cenários e identificar pontos críticos.

Em síntese, a modelagem e simulação computacional configura-se como o método mais apropriado para lidar com a complexidade dos dados, a natureza dinâmica dos processos analisados e a necessidade de proposição de melhorias com base em evidências quantitativas. Tal escolha metodológica garante o rigor, a robustez e a aplicabilidade prática dos resultados obtidos ao final da pesquisa.

4 PROPOSTA DE *FRAMEWORK* PARA CLASSIFICAÇÃO DE LICITAÇÕES

Nesta seção são descritos os métodos e procedimentos adotados para a classificação textual utilizando técnicas de *clustering*, com ênfase no algoritmo *K-Means*. O objetivo foi identificar padrões em processos licitatórios, por meio de técnicas de *clustering*, associados à ocorrência de desequilíbrio econômico-financeiro nas atas de registro de preços, visando apoiar a tomada de decisão.

A implementação das etapas descritas nesta seção é realizada utilizando a linguagem Python versão 3.13, em conjunto com as seguintes bibliotecas para processamento de linguagem natural e aprendizado de máquina:

- *Natural Language Toolkit* (NTLK): Para pré-processamento textual.
- *Scikit-learn*: Para *clustering* e validação.
- Pandas: Para manipulação e análise de dados.
- Matplotlib: Para visualização dos resultados.

Para o pré-processamento das informações da presente pesquisa, foi utilizada a biblioteca Python *Natural Language Toolkit (*NLTK), que, segundo Wang e Hu (2021), é amplamente utilizada no processamento de linguagem natural, realizando a ação de forma fácil e rápida, mas também tendo a capacidade possibilidade de ser aplicada a projetos avançados, conforme demonstrado por Loper e Bird (2002).

4.1 COLETA E PREPARAÇÃO DOS DADOS

4.1.1 Coleta de dados

A primeira etapa consistiu na coleta de dados e documentos referentes aos processos licitatórios e às atas de registro de preços. As fontes de consulta aos dados utilizados na pesquisa incluem:

- Portal de Transparência do Governo Federal: Site governamental que disponibiliza dados sobre despesas públicas e contratos, tais como empenhos efetivamente registrados, com possibilidade de filtros pelos dados do contratado ou do contratante.

- Sítio Oficial da Instituição: Sítio eletrônico da Instituição onde são disponibilizados links de acesso a documentos diversos relacionados às contratações e aquisições realizadas pelo órgão.
- Sistemas Internos: Sistemas de controle da Instituição, tais como o Sistema Eletrônico de Informações (SEI), que registram informações detalhadas sobre licitações e contratos e que podem ser acessados por usuários externos ou via pedidos de informações em portais de transparência (Fala.br).

Foram coletados dados de processos de reequilíbrio econômico-financeiro, editais, Atas de Registro de Preço e Ordens de Fornecimento.

4.1.2 Pré-processamento dos dados

A etapa inicial da pesquisa consistiu na coleta de informações a respeito de processos de pedidos de reequilíbrio econômico-financeiro autuados na Instituição no período de setembro de 2020 a setembro de 2022. O objetivo inicial foi rotular tais processos e contabilizar o número de ocorrências no período.

Após a identificação dos processos, as informações sobre os mesmos foram consolidadas em uma planilha com identificação de diversos atributos de cada processo licitatório em que houve pedido de reequilíbrio, tais como:

- Tipo da Contratação: Identificação se a licitação é referente à contratação de serviço, aquisição de material de consumo ou aquisição de material permanente.
- Vigência da Ata de Registro de Preços: Identificação do prazo de vigência da Ata de Registro de Preços da contratação. Tendo em vista que a pesquisa foi realizada com contratações embasadas tanto pela Lei n.º 8.666/1993 quanto pela Lei n.º 14.133/2021, seria possível a ocorrência de atas com vigência inferiores a doze meses.
- Há Pedido Mínimo? Em determinadas contratações, especialmente de baixo valor, os artefatos da contratação podem estipular um quantitativo mínimo a ser solicitado em cada requisição, de forma a tornar o fornecimento mais atrativo aos licitantes.

Um total de dezoito atributos, listados de forma detalhada no Quadro 3, foram elencados dentro dos processos de reequilíbrio identificados no período, visando possibilitar um primeiro vislumbre de características comuns entre os certames.

O pré-processamento dos arquivos é fundamental para a pesquisa, visto que o uso da classificação textual sem devido tratamento prévio poderia resultar em um vetor com informações desnecessárias e/ou repetidas. Este pré-processamento dos dados visa torná-los adequados às etapas posteriores.

4.2 REPRESENTAÇÃO DOS DADOS TEXTUAIS

Em continuidade à etapa de pré-processamento dos dados, realiza-se a conversão dos documentos para uma representação numérica, objetivando permitir que os algoritmos sejam capazes de identificar possíveis padrões de semelhança entre os dados analisados. Caseli e Nunes (2024) esclarecem que os modelos de representação vetorial possuem duas classificações principais, podendo ser de vetores esparsos e vetores densos. Vetores esparsos possuem como característica uma alta dimensionalidade e uma alta quantidade de valores nulos, sendo comuns em abordagens como TF-IDF e PMI. Os vetores esparsos se mostram úteis em tarefas de recuperação de informações e em análises de frequência, porém, também apresentam limitações quanto à sua capacidade de capturar relações mais profundas de semântica entre os termos, tratando cada palavra como um item isolado, não considerando o contexto geral em que ela ocorre.

Já os modelos de vetores densos, conhecidos como *embeddings*, representam palavras em um espaço dimensional menor, com valores que codificam as propriedades semânticas e contextuais a partir dos dados analisados. Esses tipos de modelo se mostram mais adequados para tarefas que envolvem aprendizados não supervisionados, a exemplo do *clustering*, uma vez que permitem identificar relações de similaridade semântica de maneira mais precisa e eficiente (Caseli; Nunes, 2024).

Considerando o contexto da presente pesquisa, o modelo *Word2Vec* foi escolhido por se tratar de um modelo de vetores densos, estáticos e com capacidade de detectar regularidades linguísticas e relações semânticas. Como demonstrado por Mikolov *et al*. (2013), este modelo aprende representações vetoriais que visam

posicionar palavras com relacionamento semântico em regiões mais próximas em espaço vetorial contínuo, favorecendo de forma significativa a fase de agrupamento dos documentos licitatórios analisados, tendo como base a similaridade textual dos mesmos. Ademais, por gerar *embeddings* estáticos, o *Word2Vec* possibilita uma maior estabilidade na representação dos termos, o que se mostra como uma característica fundamental na presente pesquisa, visto que os documentos utilizados na análise possuem um padrão jurídico normativo com pouca margem de modificação. Desta forma, a escolha do *Word2Vec* se mostra alinhada às necessidades metodológicas da pesquisa, possibilitando a detecção de padrões de semântica consistentes e a formação de *clusters* de forma mais coerente, visando a identificação de possíveis ocorrências de desequilíbrios econômico-financeiros.

4.3 APLICAÇÃO DO ALGORITMO *K-MEANS*

4.3.1 Definição do Número de *Clusters*

Uma das desvantagens inerentes ao algoritmo *K-Means* é a necessidade de pré-especificar o número de *clusters*, o parâmetro *k*. Uma escolha inadequada de *k* pode levar a um agrupamento que não representa a estrutura natural dos dados, seja agrupando *clusters* distintos (se *k* for muito baixo) ou dividindo *clusters* coesos (se *k* for muito alto). Para abordar este desafio, foi empregado o Método do Cotovelo (*Elbow Method*), uma heurística amplamente utilizada para estimar um valor apropriado para *k*.

O método consiste em executar o algoritmo *K-Means* para um intervalo de valores de *k* (por exemplo, de 1 a 10) e, para cada execução, calcular a SSE (Soma das Distâncias Quadráticas), também conhecida como inércia. A inércia mede a compactação dos *clusters*, e seu valor sempre diminuirá à medida que k aumenta, pois, mais centroides significam que os pontos estarão, em média, mais próximos do centroide mais próximo. No entanto, a taxa de diminuição da inércia tende a ser acentuada no início e depois desacelera. Ao plotar a inércia em função de *k*, o gráfico resultante geralmente se assemelha a um braço humano. O "cotovelo" da curva, o ponto onde a taxa de diminuição da inércia se torna marcadamente menos pronunciada, é considerado um indicador de um bom equilíbrio entre a compactação

dos *clusters* e o número de *clusters* utilizados. Para o conjunto de dados da UFMT, a análise da curva de inércia sugeriu que um valor de *k* em torno de 3 a 6 seria um ponto de partida razoável, indicando a presença de algumas categorias principais e distintas de documentos.

Em seção posterior será explorada a lógica utilizada para definição do número de *clusters*.

4.4 INTERPRETAÇÃO DOS *CLUSTERS* E APLICAÇÃO DE MODELOS PREDITIVOS

Nesta seção, são detalhados os processos realizados após a etapa de agrupamento dos *clusters*, especificamente a de análise dos mesmos e posteriormente a etapa seguinte, de aplicação de modelos preditivos.

4.4.1 Análise dos *Clusters*

Após a execução do processo de *clustering*, foi realizada a análise interpretativa dos grupos formados, com o intuito de identificar padrões e características comuns entre os documentos que integram cada *cluster*. Esta etapa de análise foi essencial para que fosse possível compreender as relações entre os textos analisados e para avaliar de que forma determinados atributos linguísticos e contextuais poderiam estar associados à ocorrência de desequilíbrios nos documentos analisados.

Essa análise qualitativa e quantitativa dos *clusters* formados colaborou para a identificação de grupos com maior probabilidade de risco, permitindo que a etapa seguinte, de aplicação de modelos preditivos, pudesse ser corretamente executada.

4.4.2 Uso de Modelos Preditivos

Após a identificação e interpretação dos *clusters* nos dados textuais referentes aos processos licitatórios, o próximo passo envolveu o desenvolvimento de modelos preditivos para o agrupamento de novos documentos em relação à probabilidade de ocorrência de desequilíbrios econômico-financeiros. Essa abordagem permitirá

automatizar a identificação de padrões de risco em futuros processos, facilitando a tomada de decisões preventivas.

A aplicação de modelos preditivos após a análise de *clusters* resulta em uma ferramenta robusta para a antecipação de possíveis desequilíbrios em atas de registro de preços da Instituição. Essa metodologia combinou a exploração não supervisionada dos dados com a precisão dos modelos supervisionados, resultando em um algoritmo capaz de auxiliar na gestão eficiente e preventiva dos processos licitatórios.

4.5 DAS LIMITAÇÕES NO DESENVOLVIMENTO DA PESQUISA

A presente pesquisa apresenta algumas limitações que devem ser consideradas na interpretação dos resultados obtidos. Inicialmente, destaca-se a dificuldade relacionada à coleta e organização dos dados, uma vez que os documentos utilizados (editais, atas de registro de preços e ordens de fornecimento) encontram-se em formatos heterogêneos e, em muitos casos, não estruturados, exigindo etapas adicionais de tratamento e padronização.

Adicionalmente, a implementação do modelo computacional demandou conhecimentos técnicos específicos em programação e processamento de linguagem natural, o que representou um desafio ao longo do desenvolvimento da pesquisa, especialmente na fase de testes e validação dos resultados. Essas dificuldades também se estenderam à compreensão e aplicação de conceitos matemáticos associados aos algoritmos utilizados, o que exigiu aprofundamento teórico e ajustes iterativos no modelo.

No que se refere aos testes realizados, ressalta-se que a definição de parâmetros e a avaliação dos agrupamentos envolveram um processo iterativo, sujeito a decisões analíticas que podem influenciar os resultados obtidos. Além disso, como o modelo se baseia em técnicas de aprendizado não supervisionado, não há uma referência objetiva prévia para validação absoluta dos agrupamentos, o que constitui uma limitação inerente a esse tipo de abordagem.

Por fim, destaca-se que os resultados estão condicionados ao conjunto de dados analisado, o que pode limitar a generalização dos achados para outros contextos institucionais ou bases de dados com características distintas. Apesar

dessas limitações, entende-se que a pesquisa contribui ao propor uma abordagem estruturada para análise de documentos licitatórios, oferecendo subsídios relevantes para estudos futuros e aplicações práticas na gestão pública.

5 RESULTADOS E DISCUSSÕES

5.1 DEFINIÇÃO E DETALHAMENTO DOS DADOS UTILIZADOS

A aplicação de modelos de agrupamento baseados em aprendizado não supervisionado requer um volume significativo de dados, de modo a possibilitar a identificação consistente de padrões intrínsecos. Em conjuntos mais extensos, a maior densidade de informação contribui para a formação de *clusters* mais estáveis e representativos, favorece a captura da heterogeneidade estrutural dos dados, aspecto amplamente discutido na literatura de *clustering*.

Na presente pesquisa, os dados utilizados para treinamento do algoritmo foram obtidos a partir de fontes oficiais da UFMT, sendo compostos por três categorias documentais: editais de licitação, atas de registro de preços e ordens de fornecimento. Esses documentos foram selecionados por representarem diferentes etapas do ciclo das contratações públicas e por conterem informações relevantes para a análise do equilíbrio econômico-financeiro.

Do ponto de vista analítico, não se utilizam os documentos em sua forma bruta, mas sim o conteúdo textual neles presente, incluindo descrições dos objetos contratados, especificações técnicas, quantitativos, prazos, condições de execução e demais elementos que caracterizam as contratações. Esses componentes textuais são fundamentais para capturar padrões associados às características operacionais e econômicas dos processos licitatórios.

Após a etapa de extração e pré-processamento, os textos são transformados em representações vetoriais por meio de técnicas de processamento de linguagem natural, permitindo que o algoritmo de clusterização opere sobre os dados em formato numérico. Dessa forma, os agrupamentos gerados refletem similaridades semânticas entre os documentos, possibilitando a identificação de padrões relacionados à ocorrência de desequilíbrio econômico-financeiro.

Essa abordagem permite tratar os documentos como representações vetoriais em espaço de características, possibilitando a aplicação de métricas de similaridade para formação dos *clusters*.

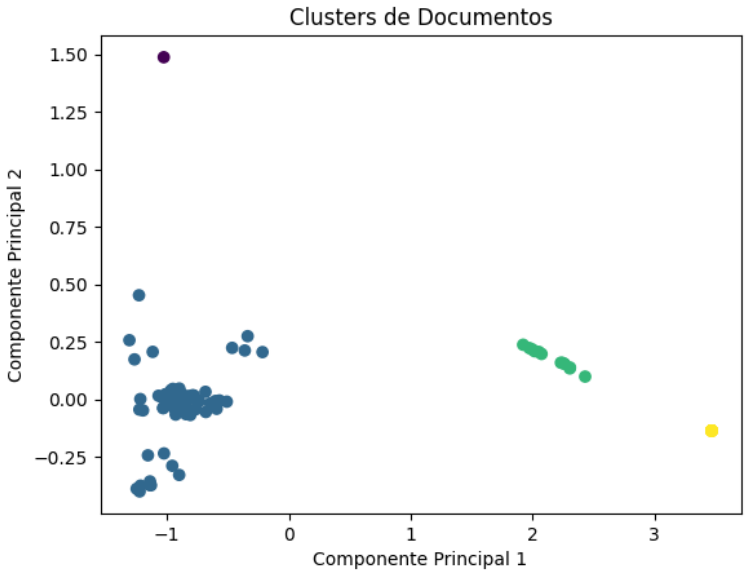
É importante salientar que o volume de documentos utilizados não se distribui de forma homogênea entre as categorias citadas. O número de OFs é

substancialmente superior ao de ARPs, que por sua vez supera o número de editais. Isso decorre em função própria dinâmica do Sistema de Registro de Preços, onde um único edital pode originar múltiplas atas, cada uma para um fornecedor distinto; por sua vez, cada ARP originada pode resultar em diversas OFs no decorrer de sua vigência, visto que cada pedido resultará em uma OF.

Os documentos utilizados passaram por um processo de padronização e consolidação, de maneira a garantir amostras íntegras e em conformidade para a aplicação do processo de classificação textual, seguindo as etapas demonstradas de maneira bastante esclarecedora no trabalho elaborado por Ikonomakis; Kotsiantis; Tampakas (2005).

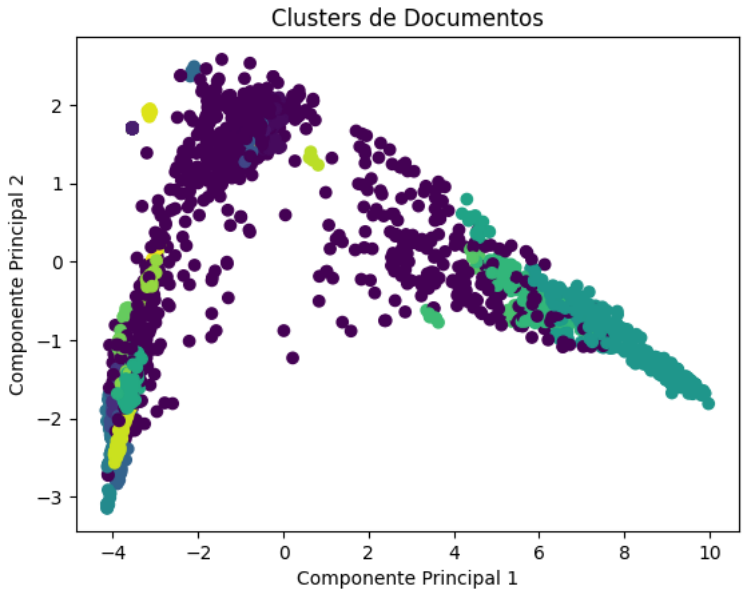
Em função da elevada exigência computacional das tarefas de processamento de linguagem natural e *clustering*, a condução dos testes foi realizada de maneira incremental, visando avaliar o desempenho do modelo proposto e o tempo de processamento da tarefa para diferentes cenários, com escalas distintas de dados. Na etapa inaugural foi utilizada uma amostra de 100 arquivos, com a tarefa sendo executada com um tempo de aproximadamente 13 minutos. Os testes foram então sendo executados com aumento gradativo dos documentos da amostra, até alcançar o total de 2.500 documentos, quando a execução de cada tarefa de processamento demandou um tempo de 2 horas e 25 minutos, o que demonstra um aumento do custo computacional, em consonância com as análises de Kaufman e Rousseuw (2005), que reforçam a relação direta entre o tamanho da amostra utilizada e a complexidade dos algoritmos de *clustering.*

Figura 6 - Testes iniciais, com 100 arquivos, 12m48s



Fonte: elaborada pelo autor (2026)

Figura 7 - Testes iniciais, com 2500 arquivos, 2h25m47s



Fonte: elaborada pelo autor (2026)

5.1.1 Dos atributos para agrupamento dos processos de reequilíbrio analisados entre 2020 e 2022

Considerando o objetivo de identificar possíveis padrões associados à ocorrência de situações de desequilíbrio econômico-financeiros em processos licitatórios que visem o registro de preços, realizou-se a construção de um conjunto

extenso de atributos que representariam as principais características de cada processo analisado, sejam administrativas, de mercado ou quanto a logística. A classificação de todo o universo de documentos utilizados para o treinamento do algoritmo seria inviável, tanto do ponto de vista metodológico quanto computacional, assim, a abordagem utilizada foi seletiva, tendo como base processos em que já havia ocorrido o desequilíbrio econômico-financeiro em licitações realizadas pela UFMT, tendo em consideração todos os pedidos recebidos no período, independente de aprovação ou não. Os pedidos de reequilíbrio utilizados para esta classificação foram autuados no período de 2020 a 2022. A amostra utilizada apresenta significativa diversidade de objetos, o que garante robustez na análise exploratória.

Os atributos apontados foram determinados de forma a permitir a avaliação e identificação de fatores com potencial para determinar a saúde econômico-financeira das contratações. A seguir, são apresentados os atributos utilizados para esta classificação:

Quadro 3 - Atributos da Classificação de Processo de Reequilíbrio

Atributo	Descrição técnica	Justificativa analítica
Tipo de material	Classifica o objeto como material permanente, material de consumo ou serviço	Reflete a natureza do bem ou serviço
Vigência da ARP	Indica a vigência da contratação, conforme legislação	ARPs com vigências maiores podem ser mais suscetíveis a variações econômicas
Exigência de garantia	Indica se há a necessidade de garantia do bem ou serviço	A presença de garantia impacta o risco do fornecedor e os custos operacionais
Objeto com validade	Determina se o bem possui prazo de validade mínimo	Relevante especialmente para bens perecíveis ou sensíveis ao tempo, como alimentos e reagentes
Objeto perecível	Classifica o bem como perecível ou não perecível	Quando o bem é perecível, aumenta o risco de perda, impactando na logística
Objeto divisível	Indica se os itens podem ser adquiridos individualmente	A divisibilidade influencia a competitividade e o número de fornecedores
Fornecedor local	Identifica se o fornecedor é sediado na mesma localidade do órgão	Impacta diretamente nos custos logísticos

Exigência de pedido mínimo	Indica se há a previsão de quantitativo mínimo por pedido	Utilizado para otimizar a logística e evitar execução fragmentada
Compra conjunta	Identifica se há participação de outros órgãos	Aumenta a escala da contratação e a complexidade logística
Licitação exclusiva ME/EPP	Define se o certame foi restrito a micro e pequenas empresas (Decreto n.º 8.538/2015)	Impacta na competitividade e no perfil dos fornecedores
Agrupamento em lotes	Indica se os itens foram agrupados em lotes	O agrupamento altera a dinâmica de disputa e a diversificação de risco
Fornecedor vencedor de múltiplos itens	Verifica se a empresa venceu mais de um item na mesma licitação	Concentração de fornecimento tende a reduzir risco de falha de execução
Fornecimento imediato (pronta entrega)	Identifica se o bem é de entrega imediata	Objetos de pronta entrega são menos suscetíveis a flutuações de mercado
Fabricação própria	Determina se o fornecedor é o fabricante do bem	A redução de intermediários pode impactar no risco econômico
Objeto de TI	Determina se o objeto é classificado como TI	Itens de TI possuem maior volatilidade e risco tecnológico
Envolve importação	Indica se há componentes importados	A importação aumenta o risco cambial e os prazos logísticos
Intervalo entre orçamento e licitação	Verifica se o intervalo entre as cotações e a fase externa da licitação foi superior a 3 meses	Longos intervalos expõem o preço ao risco inflacionário ou cambial
ARP cancelada	Informa se a ata foi cancelada após o pedido de reequilíbrio	Indicador direto de insustentabilidade econômico-contratual

Fonte: elaborada pelo autor (2026)

Os atributos elencados foram selecionados tendo por base critérios de relevância preditiva, assim como aderência normativa. Tais atributos foram utilizados como componentes que possibilitaram alimentar o modelo de *clustering*, permitindo a identificação de padrões reiterados nos processos licitatórios.

5.2 CONVERSÃO DE PDFS NÃO ESTRUTURADOS EM TEXTO PROCESSÁVEL

A primeira fase operacional do *framework* metodológico consiste na conversão dos artefatos documentais, originalmente em formato PDF, em um *corpus* textual limpo e estruturado, pronto para a etapa de vetorização. Este processo de extração e normalização é fundamental, pois a qualidade da representação de dados de entrada determina diretamente a capacidade do modelo de aprendizado de máquina de identificar padrões significativos. A negligência nesta fase pode introduzir ruídos que mascaram as relações semânticas e sintáticas subjacentes, comprometendo toda a análise subsequente.

Os primeiros passos consistiram na consolidação de um extenso banco de arquivos (editais, atas de registro de preço e ordens de fornecimento) de licitações anteriores realizadas pela Universidade Federal de Mato Grosso. Os arquivos foram então agrupados de forma a possibilitar o acesso integral via *link* direto. Em seguida, tendo em vista a utilização da ferramenta *Google Colab*, foram necessárias as instalações das bibliotecas a serem utilizadas nos diversos testes.

Cada uma das bibliotecas instaladas possui uma função específica, sendo fundamentais para os testes realizados: PyPDF2 é utilizada para manipulação de arquivos em formato PDF (leitura, extração de texto, mesclagem); nltk é utilizada no processamento de linguagem natural (tokenização e *stopwords*); gensim é um modelo de aprendizado para PLN (como *Word2Vec*); numpy para cálculos matemáticos e vetoriais; *scikit-learn* algoritmo para aprendizado de máquina (*K-Means*); e matplotlib para visualização gráfica dos dados.

A extração de conteúdo textual de arquivos PDF foi realizada utilizando a biblioteca Python *PyPDF2*. Esta biblioteca de código aberto oferece um conjunto robusto de funcionalidades para a manipulação de documentos em formato PDF, incluindo a capacidade de dividir um documento em páginas individuais, mesclar múltiplos arquivos, adicionar metadados e, mais importante para este estudo, extrair o conteúdo de texto de cada página. O processo implementado itera sobre cada arquivo PDF no diretório de dados, abrindo-o com a classe *PdfReader* e extraindo o texto de todas as suas páginas. Embora eficaz para PDFs baseados em texto, é importante reconhecer que este método pode enfrentar desafios com documentos escaneados que requerem Reconhecimento Óptico de Caracteres (do inglês, *OCR*)

ou com *layouts* complexos, o que pode resultar em erros ou perdas de informação no texto extraído.

Figura 8 - Extração de texto

```
# Função para extrair o texto de um PDF
def extract_text(pdf_file):
    pdf_reader = PyPDF2.PdfReader(pdf_file)
    text = ""
    for page_num in range(len(pdf_reader.pages)):
        page = pdf_reader.pages[page_num]
        text += page.extract_text()
    return text
```

Fonte: elaborada pelo autor (2026)

Uma vez extraído, o texto bruto passa por uma rigorosa fase de normalização linguística e pré-processamento, utilizando o *Natural Language Toolkit* (NLTK), uma das bibliotecas mais consolidadas para PLN em Python. Esta etapa envolve múltiplas operações sequenciais destinadas a reduzir a complexidade e a variabilidade do texto. A primeira operação é a tokenização, que consiste na segmentação do texto contínuo em unidades discretas, ou *tokens*, geralmente palavras ou pontuações, através da função *nltk.word_tokenize*. Este passo transforma uma cadeia de caracteres em uma lista de elementos que podem ser processados individualmente.

Figura 9 - Pré-processamento de texto

```
# Função para pré-processar o texto
def preprocess_text(text):
    words = nltk.word_tokenize(text)
    words = [word.lower() for word in words]
    words = [word for word in words if word not in stopwords.words('portuguese')]
    stemmer = PorterStemmer()
    words = [stemmer.stem(word) for word in words]
    return words
```

Fonte: elaborada pelo autor (2026)

Após a tokenização, realiza-se a remoção de *stopwords*. *Stopwords* são palavras de alta frequência que carregam pouco valor semântico para a análise de conteúdo, como artigos, preposições e conjunções (por exemplo, "de", "a", "o", "que", "e"). A sua remoção, utilizando a lista pré-definida para o português *(stopwords.words('portuguese'))* fornecida pelo NLTK, reduz a dimensionalidade do

problema e permite que o modelo se concentre nos termos que efetivamente distinguem o conteúdo dos documentos.

A etapa final do pré-processamento é a redução das palavras à sua forma radical. Existem duas abordagens principais para isso: *stemming* e lematização. O *stemming* é um processo heurístico que remove sufixos das palavras para reduzi-las a um radical comum, sendo computacionalmente rápido, mas podendo resultar em radicais que não são palavras reais (ex: "comunicação" -> "comun"). A lematização, por outro lado, é um processo mais sofisticado que utiliza análise morfológica e um dicionário para retornar a forma canônica da palavra (o lema), garantindo que o resultado seja uma palavra válida. Para este projeto, optou-se pelo *stemming*, utilizando o *PorterStemmer* do NLTK, como uma escolha pragmática que equilibra a redução de dimensionalidade com a eficiência computacional, considerando que para a tarefa de *clustering* baseada em coocorrência de termos, a preservação da validade linguística do radical é menos crítica do que a consistência no agrupamento de formas flexionadas de uma mesma palavra.

Figura 10 - Remoção de *stopwords*

```
# Remoção de stopwords
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('punkt_tab')
sentences = [preprocess_text(text) for text in texts]
```

Fonte: elaborada pelo autor (2026)

5.2.1 Transformação de texto em representações quantitativas

A aplicação de algoritmos de aprendizado de máquina para dados textuais é condicionada pela necessidade de converter a linguagem natural, intrinsecamente qualitativa, em representações numéricas. Este processo, conhecido como vetorização ou engenharia de características, mapeia palavras, frases ou documentos inteiros para vetores em um espaço matemático de alta dimensionalidade. A qualidade e a natureza dessa representação vetorial são determinantes para a capacidade do modelo de capturar as nuances semânticas e as relações entre os textos. Diversos

modelos foram desenvolvidos para essa finalidade, variando em complexidade e na forma como representam o significado.

O cálculo é um produto de duas métricas: a Frequência do Termo (do inglês, TF), que mede a frequência com que uma palavra aparece em um determinado documento, e a Frequência Inversa do Documento (do inglês, IDF), que penaliza palavras que são muito comuns em todos os documentos (como "licitação" ou "contrato" no corpus da pesquisa), atribuindo maior peso a termos que são mais distintivos de um documento específico. A fórmula do TF-IDF é dada por TF−IDF(t,d) = TF(t,d)×IDF(t), onde TF(t,d) é a frequência do termo t no documento d, e IDF(t)=log(dftN), sendo N o número total de documentos e dft o número de documentos que contêm o termo t.

Uma abordagem mais moderna e semanticamente rica é a utilização de *word embeddings*, que são representações vetoriais densas e de baixa dimensionalidade (tipicamente entre 100 e 300 dimensões) para cada palavra no vocabulário. Diferente do TF-IDF, que gera vetores esparsos e de alta dimensionalidade, os *embeddings* são aprendidos por redes neurais e têm a propriedade notável de capturar relações semânticas e sintáticas. Neste espaço vetorial, palavras com significados semelhantes são posicionadas próximas umas das outras, permitindo a realização de operações aritméticas que refletem analogias linguísticas, como a famosa expressão vetor(′Rei′)−vetor(′Homem′)+vetor(′Mulher′)≈vetor(′Rainha′).

Para a presente pesquisa, a escolha recaiu sobre a utilização de *word embeddings* devido à sua capacidade superior de capturar o significado contextual e as nuances semânticas dos termos jurídicos e administrativos presentes nos documentos de licitação. Dentre os diversos modelos de *word embeddings*, uma análise comparativa foi realizada para justificar a seleção do *Word2Vec*. O *Global Vectors for Word Representation* (*GloVe*) é um modelo baseado em contagem que constrói sua representação a partir de uma matriz global de coocorrência de palavras, capturando estatísticas de todo o corpus. O *FastText*, uma extensão do *Word2Vec*, aprende vetores para n-gramas de caracteres, o que lhe permite gerar representações para palavras fora do vocabulário de treinamento (OOV) e lidar de forma mais eficaz com linguagens morfologicamente ricas. Embora abordagens como o *FastText* possam apresentar melhor desempenho em determinadas tarefas de agrupamento, optou-se pela utilização do *Word2Vec* em função de seu equilíbrio entre desempenho

e eficiência computacional, além da robustez de sua implementação na biblioteca *Gensim*. Essa escolha mostra-se adequada ao objetivo da pesquisa, ao permitir a geração de representações vetoriais consistentes para suporte à etapa de *clustering* dos documentos.

5.2.2 O modelo *Word2Vec* e as arquiteturas CBOW e *Skip-Gram*

O modelo *Word2Vec*, proposto por Mikolov *et al*. (2013), introduziu um método computacionalmente eficiente para aprender representações de palavras de alta qualidade a partir de grandes volumes de texto. A sua eficácia reside em duas arquiteturas de redes neurais: o *Continuous Bag-of-Words* (CBOW) e o *Skip-Gram*. Ambas as arquiteturas aprendem os vetores de palavras (*embeddings*) ao treinar a rede para realizar uma tarefa de predição contextual, mas o fazem de maneiras inversas.

A arquitetura CBOW tem como objetivo prever uma palavra-alvo a partir do seu contexto. Dado um conjunto de palavras de contexto (as palavras que aparecem antes e depois da palavra-alvo dentro de uma janela definida), o modelo CBOW tenta maximizar a probabilidade da palavra-alvo. Matematicamente, o modelo calcula a média dos vetores das palavras de contexto e utiliza este vetor médio para prever a palavra central. Esta abordagem tende a ser mais rápida de treinar e funciona bem para palavras frequentes.

Por outro lado, a arquitetura *Skip-Gram* realiza a tarefa inversa: dada uma palavra-alvo, ela tenta prever as palavras de contexto que a cercam. Para cada palavra de entrada, o modelo *Skip-Gram* gera múltiplas predições para as palavras dentro da janela de contexto. Esta abordagem é computacionalmente mais intensiva, mas demonstrou ser mais eficaz na captura de representações para palavras raras e em tarefas de analogia semântica. A função objetivo em ambos os casos envolve a maximização da probabilidade logarítmica das palavras corretas, utilizando uma camada de saída com uma função *softmax* sobre todo o vocabulário.

Um dos principais desafios computacionais do *Word2Vec* é o cálculo da função *softmax* na camada de saída, que requer a normalização sobre todos os termos do vocabulário, um processo extremamente custoso para vocabulários com milhões de palavras. Para contornar este problema, Mikolov *et al*. (2013) introduziram a técnica

de otimização chamada *Negative Sampling*. Em vez de atualizar os pesos de todos os neurônios na camada de saída, o *Negative Sampling* reformula o problema. Para cada par positivo (palavra-alvo, palavra de contexto), a técnica seleciona um pequeno número de pares negativos (a palavra-alvo com palavras aleatoriamente amostradas do vocabulário que não aparecem em seu contexto). O modelo é então treinado como uma série de problemas de classificação binária, onde o objetivo é distinguir os pares positivos dos negativos. Esta aproximação reduz drasticamente a carga computacional, tornando o treinamento em grande escala viável e eficiente.

A classe *gensim.models.Word2Vec* permite um controle granular sobre o processo de treinamento através de parâmetros chave como *vector_size* (a dimensionalidade dos vetores de palavras), *window* (o tamanho da janela de contexto), min_count (o limiar de frequência para incluir uma palavra no vocabulário) e *sg* (que define a arquitetura a ser usada: 1 para *Skip-Gram*, 0 para CBOW). Quando o parâmetro *sg* não é configurado de forma explícita, considera-se a arquitetura CBOW, sendo esta o padrão da biblioteca *Gensim*. Considerando o contexto da pesquisa, onde são utilizados documentos de licitações anteriores da UFMT, os quais possuem estrutura padronizada e linguagem técnica recorrente, a arquitetura CBOW se mostra mais apropriada, uma vez que privilegia a representação dos termos com base no seu uso contextual predominante, formando vetores com maior coerência semântica.

A utilização da *Gensim* facilitou a geração de *embeddings* de alta qualidade a partir dos dados volume de documentos da UFMT, que serviram como entrada para a fase de *clustering*.

Figura 11 - Aplicação do *Word2Vec*

```
# Criar um modelo Word2Vec
model = Word2Vec(sentences, vector_size=100, window=5, min_count=1, workers=4)

# Criar representações vetoriais
def create_vector(text):
    words = preprocess_text(text)
    vectors = []
    for word in words:
        if word in model.wv:
            vectors.append(model.wv[word])
    if vectors:
        return np.mean(vectors, axis=0)
    else:
        return np.zeros(100)  # Vetor nulo se nenhuma palavra for encontrada

vectors = [create_vector(text) for text in texts]
```

Fonte: elaborada pelo autor (2026)

5.3 APRENDIZADO NÃO SUPERVISIONADO PARA A DESCOBERTA DE PADRÕES EM DOCUMENTOS DE AQUISIÇÃO

Após a transformação dos documentos textuais em representações vetoriais de alta dimensionalidade, a próxima etapa do *framework* metodológico consiste na aplicação de algoritmos de aprendizado não supervisionado para descobrir estruturas e padrões latentes. O objetivo do *clustering*, ou agrupamento, é particionar o conjunto de vetores de documentos em grupos coesos, onde os documentos dentro de um mesmo grupo são semanticamente similares entre si e distintos dos documentos em outros grupos. A escolha do algoritmo de *clustering* é uma decisão crítica, pois diferentes métodos operam sob premissas distintas sobre a estrutura dos dados, o que pode levar a resultados de agrupamento drasticamente diferentes. Nesta seção é apresentada uma análise comparativa rigorosa dos principais paradigmas de *clustering*, a justificativa pela seleção do algoritmo *K-Means* e sua variante aprimorada, *K-Means++*, além de detalhar os fundamentos matemáticos e a implementação computacional utilizados na análise empírica.

5.3.1 Princípios do Agrupamento Baseado em Densidade e Baseado em Centróides

No universo do aprendizado não supervisionado, os algoritmos de *clustering* podem ser amplamente categorizados com base na lógica que utilizam para formar

os grupos. Duas das abordagens mais proeminentes são as baseadas em centroides e as baseadas em densidade. O algoritmo *K-Means* é o principal representante do *clustering* baseado em centroides. Sua premissa fundamental é que um *cluster* pode ser representado por um único ponto central, ou centroide, e o algoritmo busca particionar os dados em *k* grupos de tal forma que a soma das distâncias quadradas de cada ponto ao centroide de seu *cluster* seja minimizada. Uma consequência implícita desta abordagem é que o *K-Means* tende a encontrar *clusters* de formato esférico ou convexo e de tamanhos relativamente uniformes.

Em contrapartida, o *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) representa o paradigma de *clustering* baseado em densidade. Este algoritmo define *clusters* como áreas de alta densidade de pontos, separadas por áreas de baixa densidade. O DBSCAN não faz suposições sobre o formato dos *clusters*, o que lhe confere a capacidade de identificar agrupamentos de formas arbitrárias e complexas. Além disso, ele possui um mecanismo intrínseco para identificar e rotular pontos que não pertencem a nenhum *cluster* denso como ruído ou *outliers*, uma vantagem significativa em muitos conjuntos de dados do mundo real.

Apesar das vantagens teóricas do DBSCAN, especialmente sua flexibilidade de forma e robustez a ruído, sua aplicação a dados de alta dimensionalidade, como os *word embeddings* utilizados nesta pesquisa, apresenta desafios substanciais. Dentre eles, o fenômeno conhecido como a "maldição da dimensionalidade" faz com que o conceito de densidade e distância se torne menos significativo à medida que o número de dimensões aumenta. Em um espaço de 300 dimensões, por exemplo, os pontos tendem a se tornar esparsos e equidistantes uns dos outros, tornando a seleção de parâmetros significativos para o DBSCAN, como *eps* (o raio da vizinhança) e *MinPts* (o número mínimo de pontos), uma tarefa extremamente difícil e muitas vezes contraintuitiva. O *K-Means*, embora mais simples em suas premissas, é frequentemente mais robusto e interpretável em espaços de alta dimensionalidade, tornando-se uma escolha pragmática e eficaz para a análise exploratória de dados textuais vetorizados.

Quadro 4 - Análise Comparativa dos algoritmos *K-Means* e DBSCAN

Característica	***K-Means***	**DBSCAN**
Formato do *Cluster*	Esférico / Convexo	Arbitrário

Tratamento de Ruído/*Outliers*	Sensível (*outliers* podem distorcer os centroides)	Robusto (identifica e rotula ruído explicitamente)
Parâmetros Necessários	Número de *clusters* (*k*)	Raio da vizinhança (*eps*) e pontos mínimos (*MinPts*)
Sensibilidade à Alta Dimensionalidade	Menos sensível; a distância euclidiana permanece funcional	Altamente sensível (a "maldição da dimensionalidade" degrada o conceito de densidade)
Justificativa da Escolha	Mais robusto e interpretável para dados textuais de alta dimensão	A seleção de parâmetros é mais complexa e os resultados são menos estáveis em espaços de alta dimensão

Fonte: elaborado pelo autor (2026)

5.3.2 Implementação com código Python e *Scikit-learn*

A implementação do *framework* de *clustering* foi realizada em *Python*, aproveitando o ecossistema robusto de bibliotecas de ciência de dados, principalmente a *Scikit-learn*. A classe *sklearn.cluster.KMeans* oferece uma implementação eficiente e flexível do algoritmo, incorporando otimizações como a inicialização *K-Means++*. A configuração do modelo para esta pesquisa utilizou parâmetros chave para garantir a robustez e a reprodutibilidade dos resultados.

O parâmetro *n_clusters* foi determinado através de análise exploratória, conforme será detalhado em seção à frente. O parâmetro *init* foi definido como '*k-means++*', instruindo o algoritmo a utilizar o método de semeadura aprimorado para a seleção dos centroides iniciais, o que é o padrão e a prática recomendada. Para mitigar o risco de convergência para um mínimo local, o parâmetro *n_init* foi configurado para um valor como 10, o que significa que o algoritmo *K-Means* completo (semeadura e iterações) é executado 10 vezes com diferentes sementes aleatórias, e o resultado final retornado é aquele com a menor inércia. O *max_iter*, que define o número máximo de iterações para uma única execução, foi mantido em seu valor padrão de 300, suficiente para garantir a convergência na maioria dos casos. Finalmente, um *random_state* foi fixado para garantir que os resultados da semeadura e, consequentemente, do *clustering*, sejam determinísticos e reprodutíveis.

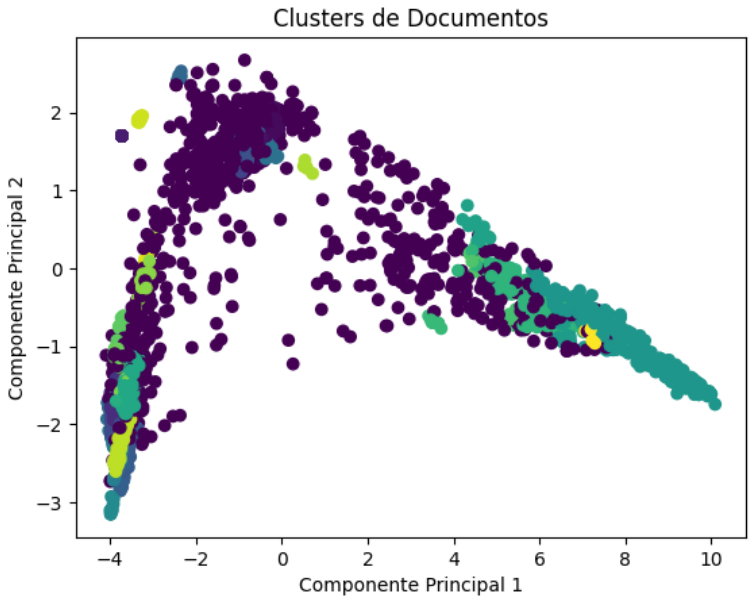
Figura 12 - Implementação do clustering

```python
from sklearn.cluster import KMeans

kmeans = KMeans(n_clusters=3, random_state=0).fit(vectors)

for i, cluster in enumerate(kmeans.labels_):
    #print(f"Cluster {i}: {pdf_files[i]}")
    #print(cluster)
    print(f"{i};{cluster};{pdf_files[i]}")

n_samples = 300
n_clusters = 3
random_state = 42
X, _ = make_blobs(n_samples=n_samples, centers=n_clusters, cluster_std=1.0, random_state=random_state)

kmeans = KMeans(n_clusters=n_clusters, init='k-means++', max_iter=300, random_state=random_state)
kmeans.fit(X)
```

Fonte: elaborada pelo autor (2026)

Uma etapa final, crucial para a interpretação dos resultados, é a visualização dos *clusters*. Como os vetores *Word2Vec* gerados possuem uma alta dimensionalidade (e.g., 300 dimensões), é impossível visualizá-los diretamente. Para superar isso, a técnica de Análise de Componentes Principais (do inglês, PCA), implementada pela classe *sklearn.decomposition.PCA*, foi utilizada. A PCA é um método de redução de dimensionalidade que projeta os dados em um subespaço de menor dimensão, preservando o máximo possível da variância original. Ao reduzir os vetores de 300 dimensões para apenas 2 componentes principais, tornou-se possível plotar cada documento como um ponto em um gráfico de dispersão 2D, com as cores dos pontos representando a atribuição de *cluster* feita pelo *K-Means*, permitindo uma inspeção visual intuitiva da separação dos grupos.

Figura 13 - Exemplo de visualização gráfica em espaço 2D



Fonte: elaborada pelo autor (2026)

5.4 ANÁLISE EMPÍRICA E INTERPRETAÇÃO DOS *CLUSTERS* DE AQUISIÇÕES

Esta seção apresenta os resultados empíricos da aplicação do *framework* de *clustering* ao corpus de documentos de licitação da UFMT. A análise transita da teoria metodológica para a descoberta de padrões concretos, oferecendo uma interpretação multifacetada das estruturas latentes nos dados. O processo analítico é apresentado de forma incremental, começando pela determinação do número ótimo de *clusters* e prosseguindo com uma exploração detalhada das configurações de agrupamento, desde uma visão macroscópica até uma granularidade mais fina. O ponto culminante desta seção é a conexão entre os *clusters* descobertos de forma não supervisionada e indicadores de risco objetivos, extraídos de metadados estruturados, validando assim a capacidade do modelo de gerar inteligência acionável para a gestão de riscos.

5.4.1 Definição do número ótimo de *clusters* com aplicação do método do cotovelo

A escolha do número adequado de *clusters* é fundamental para a eficácia do *K-Means*. Métodos como o "Método do Cotovelo *(Elbow Method)*" ou a análise do

coeficiente de silhueta podem ser empregados para determinar o valor ótimo de *clusters*.

Para a pesquisa realizada, dada a impossibilidade de definição prévia do número de *clusters* que seriam formados e pela densidade de informações analisadas, foi utilizado o Método do Cotovelo, que consiste na execução dos algoritmos com números distintos de *clusters* até que se possa encontrar o número ótimo de *clusters* (Umargono *et al*., 2019).

Para o desenvolvimento da pesquisa, o número ótimo de *clusters* encontrado nos exemplares de dados analisados foi de *k* = 3. Como detalhado anteriormente, a definição do quantitativo de *clusters* é essencial para a qualidade analítica e interpretativa dos resultados das análises realizadas. Aplicando-se o método do cotovelo, foi possível identificar o ponto de inflexão da curva SSE, a partir do qual a taxa de redução de erro se torna menos significativa com o aumento de agrupamentos.

A definição pelo número de *k* = 3, visa também favorecer a robustez estatística do modelo, de forma que os agrupamentos apresentados sejam significativos, evitando a geração de *clusters* com baixa representatividade.

Nos testes realizados, foram utilizados 3, 4, 5 e 6 *clusters*, de forma a identificar as possibilidades de melhor agrupamento dos dados. O quantitativo de documentos analisados foi estabelecido em 2.500, de forma a permitir uma ampla gama de dados, diversificados entre Atas de Registro de Preços (ARP), Ordens de Fornecimento (OF) e Editais. Os resultados obtidos demonstraram que a utilização de 4, 5 e 6 *clusters* gerou um ou mais agrupamentos com baixa representatividade no contexto geral, como demonstrado na tabela comparativa abaixo:

Quadro 5 - Comparativo de Agrupamento com *k* Variável

Número de *Clusters*	**Número do *cluster***	**Número de documentos no *cluster***	**Percentual de documentos no *cluster***
***k* = 3**	0	953	37%
	1	622	25%
	2	920	38%
***k* = 4**	0	828	33%
	1	621	25%
	2	917	37%
	3	129	5%

***k* = 5**	0	650	26%
	1	310	12%
	2	864	35%
	3	360	14%
	4	311	12%
***k* = 6**	0	524	21%
	1	312	13%
	2	865	35%
	3	129	5%
	4	303	12%
	5	362	15%

Fonte: elaborado pelo autor (2026)

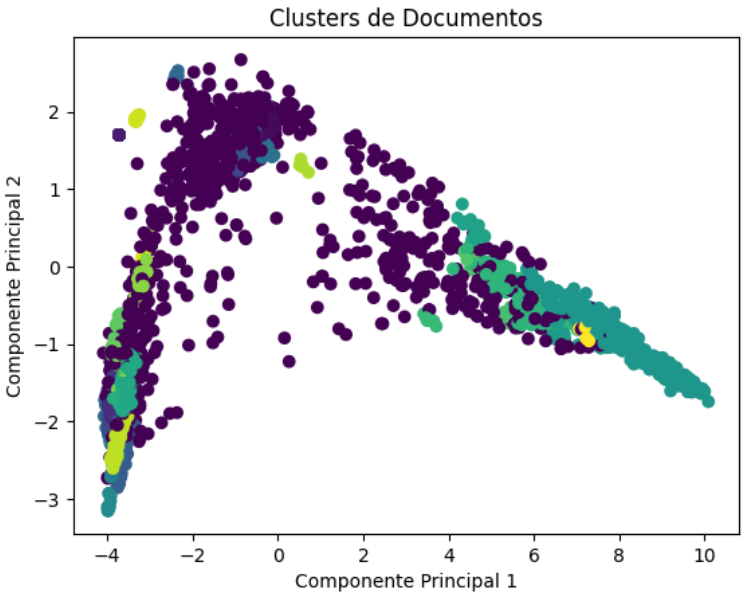
Especialmente quando definidos *k* = 4 e *k* = 6, foram gerados agrupamentos com baixo número de documentos, representando em torno de 5% do universo de documentos analisados. Quando definido *k* = 5, obtém-se 3 agrupamentos com números similares (12%, 14% e 12%, respectivamente) e outros 2 discrepantes (26% e 35%). Quando definido *k* = 3, os agrupamentos resultantes são mais próximos, demonstrando uma distribuição mais equiparada entre os *clusters*.

5.4.2 Avaliação de resultados com números variados de *clusters*

5.4.2.1 Dos resultados quando *k* = 3

Considerando o número de *clusters* ideal, a execução do algoritmo *K-Means* foi iniciada com *k* = 3, resultando em uma segmentação bastante clara e funcional da amostra documental analisada. A amostra utilizada foi composta de 2.500 documentos de licitações da UFMT (editais, ARPs e OFs), dos quais 2.495 foram efetivamente alocados entre os *clusters*, enquanto 5 documentos foram desconsiderados pelo algoritmo por não apresentarem conteúdo textual significativo após a etapa de pré-processamento. A ocorrência de registros nulos ou vazios é comum quando utilizadas bases documentais de larga escala, principalmente na existência de alguns arquivos em formato PDF que contenham páginas com imagens digitalizadas ou conteúdo ilegível, o que inviabiliza a representação vetorial dos mesmos. Tendo em vista o volume de documentos, o percentual de documentos desconsiderados (aproximadamente 0,2% do total), demonstra que o índice de efetividade da etapa de processamento textual foi superior a 99%.

Figura 14 - Visualização da análise com k = 3



Fonte: elaborado pelo autor (2026)

Quando configurado com *k* = 3, os resultados demonstram alta coerência semântica, revelando segregação dos documentos de acordo com a categoria em que os mesmos se inserem nas etapas do ciclo do processo licitatório. A interpretação de partida dos agrupamentos foi realizada com base na análise da nomenclatura dos arquivos e nas terminologias predominantes em seu conteúdo. Por esta análise, foi possível caracterizar cada *cluster* conforme sua função no ciclo da licitação.

Quadro 6 - Resumo da Composição dos *Clusters* (k = 3)

ID do *Cluster*	**Número de Documentos**	**Tipos de Documentos Predominantes**	**Descrição Qualitativa**
0	953	Edital, ARP (Ata de Registro de Preços)	Documentos da Fase de Planejamento e Fase Externa
1	622	OF (Ordem de Fornecimento)	Documentos da Fase de Execução (Grupo A)
2	920	OF (Ordem de Fornecimento), ARP (Ata de Registro de Preços)	Documentos da Fase Externa e de Execução (Grupo B)

Fonte: elaborado pelo autor (2026)

O *cluster* “0” é formado por documentos que estão associados às fases de planejamento (Editais) e de execução (ARPs). Tais documentos compartilham características de vocabulário jurídico normativo, tratando especialmente das regras

da licitação, obrigação das partes (contratante e contratada), condições gerais e critérios de julgamento, o que justifica e demonstra a coesão semântica dentro do agrupamento.

Figura 15 - Visualização do *cluster* “0”, com *k* = 3

Cluster	Arquivo
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/ARP 124-2020 - PE 49-2020 - TECNOLOGIA INFORMACA?O COMUNICACA?O PARA TODOS EIRELI.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/ARP 123-2020 - PE 49-2020 - CAM TECNOLOGIA EIRELI ME.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/ARP 125-2020 - PE 55-2020 - UHLIG & KOROVSKY TECNOLOGIA LTDA.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/ARP 122-2020 - PE 44-2020 - ISOFEN ENERGY ENGENHARIA DE SUSTENTABILIDADE LTDA.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/ARP 013-2022 - PE 35-2021 - LICITAO CONSULTORIA PROJETOS E SERVICOS LTDA - EPP.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/ARP 003-2022 - PE 29-2021 - ALBR - INDU?STRIA E COME?RCIO LTDA.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/ARP 004-2022 - PE 29-2021 - W MARCHIOLI E CIA LTDA.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/ARP 005-2022 - PE 29-2021 - LIMATEC INDU?STRIA E SERVIC?OS EIRELI - EPP.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/ARP 006-2022 - PE 29-2021 - DOMI?NIO COME?RCIO DE EQUIPAMENTOS EIRELI - EPP.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/ARP 007-2022 - PE 29-2021 - PONTO COMPRAS LTDA.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/Edital 03-2022_com data.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/Edital 02-2022 -PE TRAD- Contratac?a?o de Ma?o de Obra - Apoio Administrativo - Po?s Questionamentos 03-03-2022.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/ARP 008-2022 - PE 34-2021 - LAERDAL MEDICAL IMPORTACAO E COMERCIO DE PRODUTOS MEDICOS LTDA.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/ARP 009-2022 - PE 34-2021 - WEBLABOR SA?O PAULO MATERIAIS DIDA?TICOS LTDA.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/ARP 010-2022 - PE 34-2021 - CONSULAB DISTRIBUIDORA DE PRODUTOS LABORATORIAIS, HOSPITALARES E EDUCACIONAIS – LTDA - ME.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/ARP 011-2022 - PE 34-2021 - ESFERA MASTER COMERCIAL EIRELI - EPP.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/ARP 012-2022 - PE 34-2021 - EGR EQUIPAMENTOS E SOLUC?O?ES EIRELI.pdf
0	/content/drive/MyDrive/1. Mestrado/Testes de Codigo/Material teste/Parcelas/Edital 01-2022 -SRP- COM DATA - Aquisic?a?o Eventual de Material Permanente - Termocicladores.pdf

Fonte: elaborada pelo autor (2026)

Já os *clusters* “1” e “2” são formados por documentos relacionados especialmente à fase de execução contratual, alocando os documentos identificados como Ordens de Fornecimento (OF). A criação de *clusters* distintos formados por documentos aparentemente de mesma natureza demonstra a capacidade do modelo em detectar distinções mais sutis, possivelmente associadas com características como períodos de execução, perfil de fornecedores, tipos distintos de objeto, entre outras. Ademais, enquanto o *cluster* “1”, é formado integralmente por documentos do tipo “Ordem de Fornecimento”, o *cluster* “2”, apesar de ser predominantemente formado por OFs, possui um pequeno número de ARPs agrupados, reforçando o poder de detecção de características sutis entre os documentos.

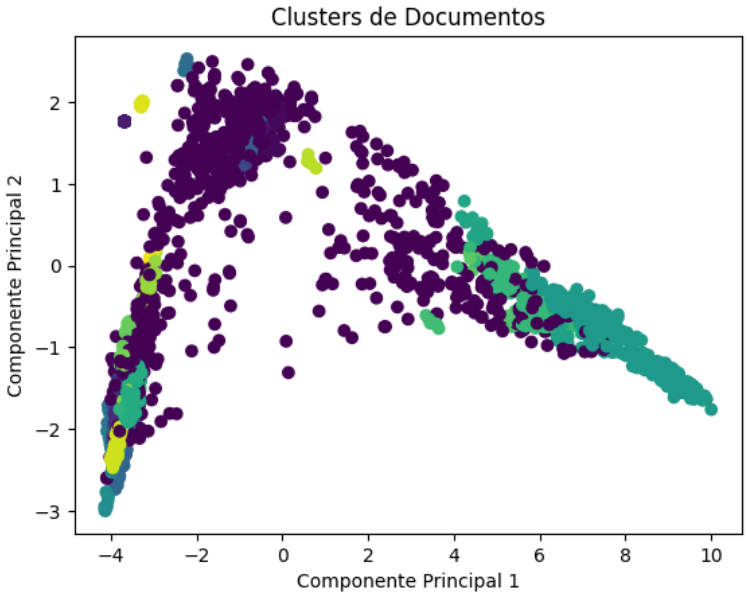
A separação percebida, demonstra a eficácia do modelo em detectar distinções latentes no conteúdo dos documentos analisados, reiterando a premissa de que os arquivos possuem estrutura semântica complexa.

5.4.2.2 Dos resultados quando *k* = 4, *k* = 5 e *k* = 6

O aumento do número de *clusters* não promove uma distribuição arbitrária dos dados analisados, mas demonstra uma estruturação hierárquica latente da amostra documental analisada. A análise dos resultados obtidos para *k* = 4, *k* = 5 e *k* = 6 evidencia que os agrupamentos realizados com *k* = 3 são decompostos de forma gradual em grupos menores e semanticamente semelhantes, refletindo possíveis

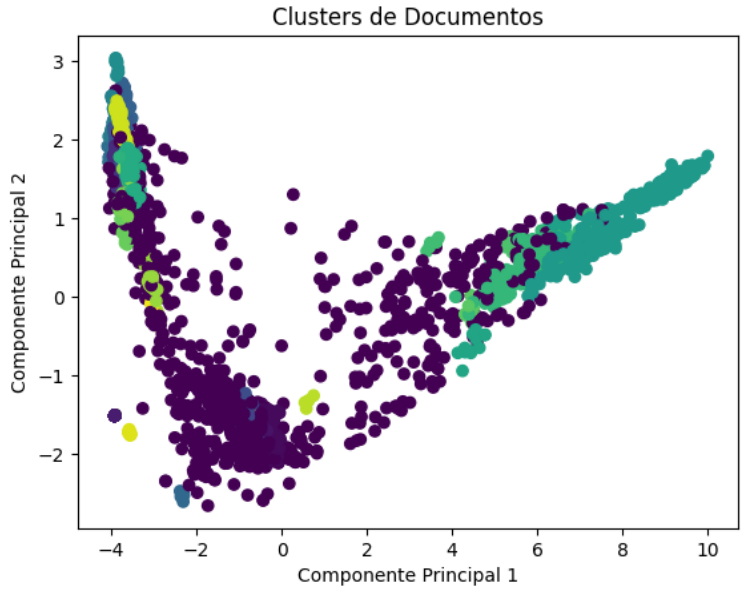
subdivisões temáticas, procedimentais e temporais dos dados analisados. O comportamento do modelo demonstra que os vetores gerados possuem a capacidade de detectar nuances semânticas significativas, permitindo que o algoritmo *K-Means* tenha condições para segregar o espaço vetorial com base em eixos inerentes ao conteúdo textual.

Figura 16 - Visualização da análise com $k = 4$



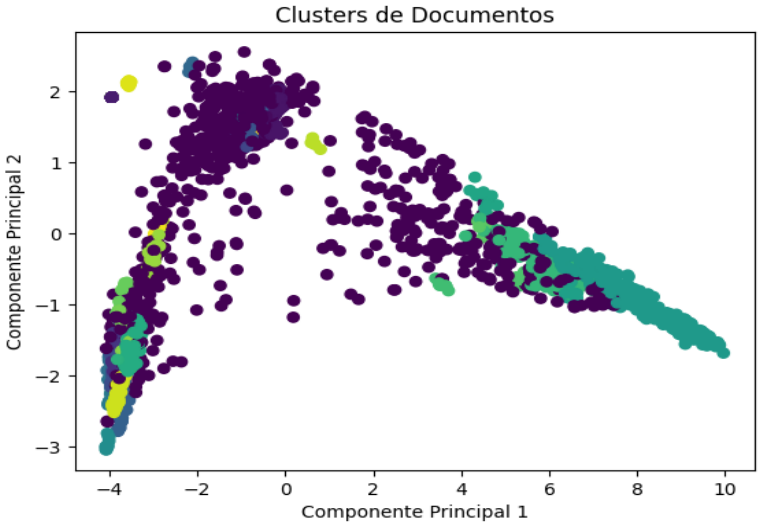
Fonte: elaborada pelo autor (2026)

Figura 17 - Visualização da análise com $k = 5$



Fonte: elaborada pelo autor (2026)

Figura 18 - Visualização da análise com *k* = 6



Fonte: elaborada pelo autor (2026)

Na tabela 7 abaixo, verifica-se que o *cluster* "0", que com *k* = 3 agrupou grande parte dos documentos associados à fase de planejamento, quando definido *k* = 6 é subdividido de forma mais granular, permitindo a detecção de padrões que não eram perceptíveis na estrutura inicial, com *k* = 3. O *cluster* foi desmembrado em três agrupamentos distintos (*clusters* "0", "3" e "4"), com cada um representando nuances específicas especialmente relacionadas ao tipo de documento (todos os documentos do tipo "Edital" foram agrupados no *cluster* "0", por exemplo) e ao período temporal de emissão.

Quadro 7 - Resumo da Composição dos *Clusters* (*k* = 6)

ID do *Cluster*	**Número de Documentos**	**Tipos de Documentos Predominantes**	**Descrição Qualitativa**
0	524	Edital	Documentos da Fase de Planejamento
1	312	OF (2017, 2018, 2019 e 2020)	Documentos da Fase de Execução (Grupo A)
2	865	OF (2021, 2022, 2023 e 2024)	Documentos da Fase de Execução (Grupo B)
3	129	OF (2017 e 2018), ARP (2020 e 2022)	Documentos da Fase de Planejamento e Execução (Grupo A)
4	303	ARP (2020, 2021, 2022 e 2023)	Documentos da Fase execução (Grupo B)
5	362	OF (2017, 2019, 2020 e 2021)	Documentos da Fase de Execução (Grupo C)

Fonte: elaborado pelo autor (2026)

O agrupamento dos documentos do tipo “Edital” no *cluster* “0”, demonstra que o modelo foi capaz de identificar as semelhanças semânticas dos documentos em função da linguagem técnica característica dos instrumentos convocatórios, os quais possuem terminologias recorrentes como “cláusulas”, “habilitação”, “exigências”, entre outras. O foco deste *cluster* está no núcleo normativo do ciclo licitatório.

Já na formação dos *clusters* “3” e “4”, há uma segmentação mais rebuscada, com a distribuição de documentos da parte final da fase externa (ARPs) e da fase de execução pelo eixo temporal e pela natureza do procedimento. O *cluster* “4” é formado por ARPs dos anos de 2020 a 2023, indicando uma formação voltada para a etapa final da fase externa. Quanto ao *cluster* “3”, este é formado tanto por documentos do tipo “ARP” quanto do tipo “OF”, mas com um número reduzido de documentos agrupados (129 ou 5,17% do total de documentos).

Os *clusters* “1”, “2” e “5”, são claramente agrupados em função do eixo temporal dos documentos analisados. O *cluster* “1” agrupa Ordens de Fornecimento (OF) com data de emissão anteriores à entrada em vigor da Lei n.º 14.133/2021 (2017 a 2020), o que indica uma proximidade semântica entre os documentos. O *cluster* “2” agrupa OFs emitidas no período de 2021 a 2024, o que reflete as atualizações normativas percebidas nos artefatos posteriores à nova lei de licitações. Já o *cluster* “5” traz um agrupamento variado quanto ao eixo temporal, mas com objetos associados a materiais de consumo.

5.4.3 Análise dos resultados e potencialidade preditiva do modelo

A execução do modelo de *clustering* não supervisionado demonstrou potencial para identificar padrões contextuais e textuais relevantes nos documentos analisados, evidenciando uma relação consistente entre padrões identificados nos documentos e a ocorrência de desequilíbrio econômico-financeiro. Os resultados indicam que tais ocorrências tendem a se concentrar em *clusters* específicos, permitindo a identificação de grupos com maior propensão a apresentar instabilidades contratuais.

Como vemos na tabela a seguir, embora os resultados não sejam de concentração de todos os documentos com registro de ocorrência de desequilíbrio em um único *cluster*, os mesmos evidenciam uma concentração bastante significativa dos

casos em agrupamentos semanticamente coerentes, o que indica que os documentos com maior probabilidade de ocorrência de desequilíbrio compartilham semelhanças textuais e características específicas relacionadas à critérios múltiplos, como objeto da contratação, questões logísticas ou outras variáveis econômicas. Ademais, considerando a forma de agrupamento dos dados, percebe-se que os documentos do tipo “Edital” foram consolidados em um *cluster* exclusivo, corroborando a interpretação da capacidade do modelo em identificar a baixa relevância específica deste artefato à questão do desequilíbrio, uma vez que são utilizadas minutas padronizadas, com baixo percentual de modificações. Os artefatos de maior relevância são as ARPs e OFs, as quais contém informações concretas quanto a valores, prazos de entrega, vigência, entre outras.

Na configuração com $k = 5$, por exemplo, o *cluster* “4” apresentou uma taxa de concentração dos casos de desequilíbrio de 63%, o que evidencia o mesmo com alto risco. O percentual se mostra ainda mais relevante quando tido em consideração a proporção deste agrupamento em relação aos demais (306 documentos, 12% do volume total).

A definição dos *clusters* foi realizada partindo do posicionamento vetorial. Cada arquivo foi convertido em vetor semântico e então agrupado pelo algoritmo com base em sua semelhança contextual. Assim, não houve definição de quaisquer critérios manuais ou temáticos, sendo que os agrupamentos emergiram a partir das relações linguísticas identificadas pelo modelo.

Quadro 8 - Análise da Composição dos *Clusters* ($k = 5$) e Percentual de Concentração de Reequilíbrio

ID do *Cluster*	**Nº de Docs.**	**Tipos Predominantes**	**Período Dominante**	**Interpretação Qualitativa**	**Percentual de Reequilíbrios anteriores**
0	525	Edital	-	Editais foram agrupados em um único *cluster*	11,21%
1	621	OF, NE	2017 e 2018	Insumos de laboratório e bens de consumo	0%
2	914	OF, NE	2021 a 2024	Aquisições de TI e outras pós	0,93%

				Lei n.º 14.133/2021	
3	129	OF, ARP	2018, 2020 e 2022	Equipamentos e bens permanentes	26,17%
4	306	ARP	2020 a 2023	Aquisições impactadas por variação mercadológica (COVID-19)	62,62%

Fonte: elaborado pelo autor (2026)

Ainda tendo como exemplo a configuração com *k* = 5, percebe-se que a distribuição dos documentos entre os grupos demonstra uma coerência estrutural na segmentação. No *cluster* “0” foram agrupados todos os editais, o que demonstra capacidade de identificação das semelhanças textuais; os *clusters* “1” e “2” concentram documentos do tipo ordem de fornecimento e nota de empenho, que representam as fases de execução, mas com distinção temporal entre os mesmos; por fim, os *clusters* “3” e “4” apresentam um baixo volume de arquivos (435, ou 17% do volume total), mas concentram aproximadamente 89% das ocorrências de desequilíbrio.

5.5 IMPLICAÇÕES ESTRATÉGICAS E APLICAÇÕES PARA A GESTÃO DO SETOR PÚBLICO

A transição da análise de dados para a geração de valor estratégico é o objetivo final de qualquer iniciativa de *data science* aplicada à gestão pública. Os *clusters* e padrões identificados na seção anterior não são meramente descrições acadêmicas da estrutura do corpus documental; eles representam uma fonte de inteligência acionável que pode transformar fundamentalmente a maneira como a administração pública planeja, executa e monitora seus processos de aquisição. Esta seção final sintetiza os achados empíricos e os projeta em um *framework* de aplicações inovadoras, propondo um sistema concreto e *data-driven* para aprimorar a gestão de compras na UFMT e, potencialmente, em outras instituições públicas, alinhando-se diretamente com a demanda do usuário por soluções inovadoras e coerentes com a realidade.

5.5.1 Inteligência acionável na gestão de aquisições

A interpretação estratégica dos *clusters* identificados permite traduzir os padrões textuais em diagnósticos operacionais. Cada *cluster* de alto risco funciona como um arquétipo de um processo de aquisição problemático, cujas características comuns fornecem pistas sobre as causas raiz das falhas. Por exemplo, um *cluster* com alta taxa de cancelamento de ARPs, composto por licitações de *software* com especificações técnicas complexas e longos prazos de entrega, pode indicar que os termos de referência estão sendo mal elaborados, que a pesquisa de mercado é insuficiente para capturar a volatilidade dos preços de tecnologia, ou que os fornecedores qualificados para tais projetos são poucos e avessos a contratos de longo prazo em um mercado dinâmico.

Esta inteligência acionável permite que os gestores de aquisições passem de uma postura reativa para uma proativa. Em vez de lidar com os problemas apenas quando eles se materializam como um pedido de reequilíbrio ou a necessidade de cancelar uma ata, a gestão pode intervir preventivamente. Os *insights* derivados dos *clusters* podem ser usados para aprimorar o planejamento das contratações, por exemplo, revisando e padronizando os termos de referência para as categorias de aquisição identificadas como de alto risco. Podem também direcionar os esforços de auditoria e controle interno, focando a atenção dos órgãos de fiscalização nos processos que se enquadram nos arquétipos mais problemáticos, otimizando assim o uso de recursos de supervisão.

5.5.2 Sistema preditivo para a mitigação do desequilíbrio econômico-financeiro

Com base nos resultados da análise de *clustering*, é possível projetar a arquitetura de um sistema de suporte à decisão para a gestão de licitações. Este sistema inovador, que atende diretamente ao objetivo central da pesquisa, funcionaria como um mecanismo de alerta precoce, integrando-se ao fluxo de trabalho da equipe de aquisições. A arquitetura proposta para este sistema preditivo seguiria um fluxo operacional claro:

- Entrada de Dados: Quando um novo processo de licitação é iniciado e um rascunho do termo de referência ou edital é elaborado, o documento de texto seria submetido ao sistema.
- Processamento e Vetorização: O sistema executará automaticamente o *pipeline* de PLN, realizando a extração de texto, a limpeza, a normalização e a conversão do documento em um vetor *Word2Vec*.
- Agrupamento por *Cluster*: O vetor do novo documento seria então passado para o modelo *K-Means* pré-treinado, que, através do método *predict()*, o atribuiria instantaneamente ao *cluster* mais próximo no espaço vetorial.
- Geração de Escore de Risco e Alerta: O sistema consultaria uma base de dados de perfis de risco dos *clusters*, que conteria informações como a taxa histórica de cancelamento de ARPs, a frequência de pedidos de reequilíbrio e outros indicadores de instabilidade para cada *cluster*. Com base no *cluster* ao qual o novo edital foi atribuído, o sistema geraria um escore de risco (e.g., baixo, médio, alto).
- Recomendação Acionável: Se o escore de risco ultrapassar um determinado limiar, o sistema emitirá um alerta para os gestores responsáveis. O alerta não seria apenas uma notificação, mas viria acompanhado de um relatório detalhado sobre as características do *cluster* de alto risco, os problemas historicamente associados a ele e um conjunto de recomendações, como "sugere-se revisão aprofundada das especificações técnicas por um especialista da área" ou "recomenda-se realizar uma nova e mais ampla pesquisa de mercado devido à alta volatilidade de preços nesta categoria".

Este *framework* transforma a análise de dados em uma ferramenta de gestão dinâmica, capacitando a UFMT a possibilitar um monitoramento mais diligente em relação a licitações sinalizadas e a mitigar proativamente os riscos de desequilíbrio econômico-financeiro antes que eles comprometam os recursos públicos e a continuidade dos serviços.

6 CONSIDERAÇÕES FINAIS

O *framework* desenvolvido nesta pesquisa estabelece uma base sólida, mas também abre um vasto campo para futuras expansões e aprimoramentos. A fronteira analítica pode ser empurrada em várias direções promissoras, aumentando a precisão, o escopo e a sofisticação do modelo de risco.

Uma primeira direção é a adoção de modelos de PLN mais avançados. Embora o *Word2Vec* tenha se mostrado eficaz, ele gera *embeddings* estáticos, ou seja, cada palavra tem um único vetor, independentemente do contexto. Modelos baseados na arquitetura *Transformer*, como o *Bidirectional Encoder Representations from Transformers* (BERT), geram *embeddings* contextuais, o que significa que o vetor de uma palavra muda dependendo da frase em que ela aparece. A utilização de um modelo como o BERT poderia capturar com ainda mais nuances as sutilezas da linguagem jurídica e técnica dos documentos de licitação, potencialmente levando a *clusters* mais precisos e informativos.

Uma segunda via de evolução é a transição para o aprendizado supervisionado. Os rótulos de *cluster* gerados pelo *K-Means*, juntamente com os indicadores de risco dos metadados (como CANCELAMENTO ARP?), podem ser usados como dados de treinamento para um modelo de agrupamento supervisionado (e.g., Regressão Logística, *Gradient Boosting*, ou uma rede neural). Tal modelo seria treinado para prever diretamente a probabilidade de um novo documento resultar em um evento adverso, oferecendo um escore de risco mais granular e direto do que a atribuição a um *cluster*.

Finalmente, a escalabilidade e a generalização do modelo representam uma fronteira importante. O *framework* poderia ser expandido para analisar dados de outras universidades federais ou mesmo de um escopo nacional, utilizando fontes como o Portal de Transparência do Governo Federal. Um modelo treinado em um conjunto de dados tão vasto e diversificado teria o potencial de descobrir padrões de risco universais na contratação pública brasileira. Além disso, a integração de fontes de dados adicionais, como séries temporais de índices de preços de insumos, histórico de desempenho de fornecedores e dados detalhados da execução contratual, poderia enriquecer o modelo, permitindo uma análise ainda mais

sofisticada e preditiva, consolidando a ciência de dados como um pilar central da governança pública moderna.

REFERÊNCIAS

AGGARWAL, C. C. **Machine Learning for Text**. Cham: Springer, 2018.

AGUIAR, F. H. O.; SAMPAIO, M. Identificação dos fatores que afetam a ruptura de estoque utilizando análise de agrupamentos. **Production**, São Carlos, v. 24, n. 1, p. 57-70, 2014. Disponível em: https://www.scielo.br/j/prod/a/nZjCfXvmHn65mJXXBrSCgps/?lang=pt. Acesso em: 2 nov. 2025.

ALLAHYARI, M. *et al*. **A brief survey of text mining: classification, clustering and extraction techniques**. 2017. Disponível em: https://doi.org/10.48550/arXiv.1707.02919. Acesso em: 2 nov. 2025.

ALVES, R. **Filosofia da ciência:** introdução ao jogo e suas regras. São Paulo: Edições Loyola, 2004.

ARBELAITZ, O. *et al*. An extensive comparative study of cluster validity indices. **Pattern Recognition**, v. 46, n. 1, p. 243-256, 2013. Disponível em: https://doi.org/10.1016/j.patcog.2012.07.021. Acesso em: 2 nov. 2025.

ARTHUR, D.; VASSILVITSKII, S. K-means++: the advantages of careful seeding. *In*: **ACM-SIAM SYMPOSIUM ON DISCRETE ALGORITHMS** (SODA), 18., 2007, New Orleans. Philadelphia: SIAM, 2007. p. 1027–1035.

BERTRAND, J. W. M.; FRANSOO, J. C. Operations management research methodologies using quantitative modeling. **International Journal of Operations & Production Management**, v. 22, n. 2, p. 241-264, 2002. Disponível em: https://doi.org/10.1108/01443570210414338. Acesso em: 2 nov. 2025.

BIRD, S.; LOPER, E. NLTK: The Natural Language Toolkit. *In*: **ACL-02 WORKSHOP ON EFFECTIVE TOOLS AND METHODOLOGIES FOR TEACHING NATURAL LANGUAGE PROCESSING AND COMPUTATIONAL LINGUISTIC**S, 2002, Philadelphia. Philadelphia: ACL, 2002. p. 63–70. Disponível em: https://aclanthology.org/W02-0109/. Acesso em: 2 nov. 2025.

BRASIL. **Constituição da República Federativa do Brasil de 1988**. Brasília, DF: Senado Federal, 1988. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 2 nov. 2025.

BRASIL. **Decreto nº 11.462, de 31 de março de 2023**. Brasília, DF: Presidência da República, 2023. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2023/decreto/d11462.htm. Acesso em: 2 nov. 2025.

BRASIL. **Emenda Constitucional nº 95**, **de 15 de dezembro de 2016**. Brasília, DF: Presidência da República, 2016. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/emendas/emc/emc95.htm. Acesso em: 2 nov. 2025.

BRASIL. **Lei nº 8.666, de 21 de junho de 1993**. Brasília, DF: Presidência da República, 1993. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l8666cons.htm. Acesso em: 2 nov. 2025.

BRASIL. **Lei nº 14.133, de 1º de abril de 2021**. Brasília, DF: Presidência da República, 2021. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm. Acesso em: 2 nov. 2025.

BRYMAN, A. **Research methods and organization studies**. London: Routledge, 2015.

CAETANO, E. F. S.; CAMPOS, I. M. B. M. A autonomia das universidades federais na execução das receitas próprias. **Revista Brasileira de Educação**, v. 24, e240043, 2019. Disponível em: https://doi.org/10.1590/S1413-24782019240043. Acesso em: 2 nov. 2025.

CASELI, H. de M.; NUNES, M. das G. V. **Processamento de linguagem natural:** conceitos, técnicas e aplicações em português. 2. ed. São Carlos: BPLN, 2024. Disponível em: https://brasileiraspln.com/livro-pln/2a-edicao/Livro_PLN_BPLN_2ed.pdf. Acesso em: 2 nov. 2025.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. **Anomaly detection: A survey**. ACM Computing Surveys, v. 41, n. 3, p. 1–58, 2009.

COOMBES, C. E. *et al*. Simulation-derived best practices for clustering clinical data. **Journal of Biomedical Informatics**, v. 118, p. 103788, 2021. Disponível em: https://doi.org/10.1016/j.jbi.2021.103788. Acesso em: 2 nov. 2025.

CRESWELL, J. W. **Research design: qualitative, quantitative, and mixed methods approaches***.* 4. ed. Thousand Oaks: Sage, 2014.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 1, n. 2, p. 224-227, 1979. Disponível em: https://doi.org/10.1109/TPAMI.1979.4766909. Acesso em: 2 nov. 2025.

DEZA, M. M.; DEZA, E. **Encyclopedia of Distances**. Berlin; Heidelberg: Springer, 2009. Disponível em: https://doi.org/10.1007/978-3-642-00234-2. Acesso em: 2 nov. 2025.

EVERITT, B. *et al*. **Cluster analysis**. 5. ed. Chichester: John Wiley & Sons, 2011.

FIELD, A. **Discovering Statistics Using IBM SPSS Statistics**. 5. ed. London: Sage Publications, 2018. Disponível em: https://edge.sagepub.com/field5e. Acesso em: 2 nov. 2025.

FORGY, E. W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. **Biometrics**, v. 21, p. 768–769, 1965. Disponível em: https: //api.semanticscholar.org/CorpusID:118110564. Acesso em: 2 nov. 2025.

FREITAS, M.; MALDONADO, J. M. S. de V. O pregão eletrônico e as contratações de serviços contínuos. **Revista de Administração Pública**, v. 47, n. 5, p. 1265-1281, 2013. Disponível em: https://doi.org/10.1590/S0034-76122013000500009. Acesso em: 2 nov. 2025.

GAO, C. X. *et al*. An overview of clustering methods with guidelines for application in mental health research. **Psychiatry Research**, v. 327, p. 115265, 2023. Disponível em: https://doi.org/10.1016/j.psychres.2023.115265. Acesso em: 2 nov. 2025.

GIL, A. C**. Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

GONÇALVES, V. B. *et al*. O sistema de registro de preços: uma maior efetividade ao princípio constitucional da eficiência. In: **II Simpósio de Inovação, Empreendedorismo e Gestão Pública**, Lavras - MG, 2018. Disponível em: https://www.researchgate.net/publication/352954967_. Acesso em: 2 nov. 2025.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge, MA: MIT Press, 2016.

GRINNELL, R. M.; UNRAU, Y. A. **Social work research and evaluation:** quantitative and qualitative approaches. 7. ed. New York: Oxford University Press, 2005.

HAHSLER, M.; PIEKENBROCK, M.; DORAN, D. dbscan: Fast Density-Based Clustering with R. **Journal of Statistical Software**, v. 91, n. 1, p. 1–30, 2019. Disponível em: https://doi.org/10.18637/jss.v091.i01. Acesso em: 2 nov. 2025.

HAIR, J. F. *et al*. **Multivariate data analysis**. 8. ed. Andover: Cengage Learning, 2019.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: concepts and techniques**. 3. ed. Burlington: Elsevier, 2011.

HAMMING, R. W. Error detecting and error correcting codes. Bell System Technical Journal, v. 29, n. 2, p. 147–160, 1950. Disponível em: https://doi.org/10.1002/j.1538-7305.1950.tb00463.x. Acesso em: 2 nov. 2025.

HERDIANA, I. *et al*. **A more precise Elbow method for optimum K-means clustering**, 2025. Disponível em: https://doi.org/10.48550/arXiv.2502.00851. Acesso em: 2 nov. 2025.

IKONOMAKIS, E.; KOTSIANTIS, S.; TAMPAKAS, V. Text classification: a recent overview. **Journal of Information Science & Technology**, 2005. Disponível em: https://www.researchgate.net/publication/234812606. Acesso em: 2 nov. 2025.

INÁCIO, L. *et al*. Educação superior no Brasil: o impacto do “Teto dos Gastos” no orçamento das universidades federais. **Revista de Políticas Públicas**, v. 28, n. 1, p. 339–359, 2024. Disponível em: https://doi.org/10.18764/2178-2865.v28n1.2024.19. Acesso em: 2 nov. 2025.

JAIN, A. K. Data clustering: 50 years beyond k-means. **Pattern Recognition Letters**, v. 31, n. 8, p. 651–666, 2010. Disponível em: https://doi.org/10.1016/j.patrec.2009.09.011. Acesso em: 2 nov. 2025.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys**, v. 31, n. 3, p. 264–323, 1999. Disponível em: https://doi.org/10.1145/331499.331504. Acesso em: 2 nov. 2025.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing**. 2. ed. Upper Saddle River, NJ: Prentice Hall, 2014.

KANUNGO, T. *et al*. An efficient K-Means clustering algorithm: Analysis and implementation**. IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 24, n. 7, p. 881–892, 2002. Disponível em: https://doi.org/10.1109/TPAMI.2002.1017616. Acesso em: 2 nov. 2025.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data:** an introduction to cluster analysis. New York: John Wiley & Sons, 2005.

LAKATOS, E. M; MARCONI, M. de A. **Fundamentos de metodologia científica***.* 8. ed. São Paulo: Atlas, 2017.

LAKATOS, R. *et al*. A Machine Learning-Based Pipeline for the Extraction of Insights from Customer Reviews*.* **Big Data and Cognitive Computing**, v. 8, n. 3, p. 20, 2024. Disponível em: https://doi.org/10.3390/bdcc8030020. Acesso em: 10 nov. 2025.

LARRAÑAGA, Pedro *et al*. **Machine learning in bioinformatics. Briefings in Bioinformatics**, v. 7, n. 1, p. 86–112, 2006.

LLOYD, S. P. Least squares quantization in pcm**. IEEE Transactions on Information Theory**, v. 28, n. 2, p. 129–137, mar 1982. ISSN 1557-9654. Disponível em: https://doi.org/10.1109/TIT.1982.1056489. Acesso em: 2 nov. 2025.

MACHADO, N. R. C. *et al*. Uma análise do sistema de registro de preços perante o princípio da eficiência. **Juris Plenum Direito Administrativo**, n. 19, 2018. Disponível em: https://www.researchgate.net/publication/351671080_. Acesso em: 2 nov. 2025.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. 1967. **Proc. 5th Berkeley Symp. Math. Stat. Probab**., Univ. Calif. 1965/66, 1, p. 281-297,1967. Disponível em: https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Some-methods-for-classification-and-analysis-of-multivariate-observations/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992. Acesso em: 2 nov. 2025.

MARTINS, R. A. Abordagens quantitativa e qualitativa. *In*: MIGUEL, Paulo Augusto Cauchick (org.). **Metodologia de pesquisa em Engenharia de Produção e Gestão de Operações**. 2. ed. Rio de Janeiro: Elsevier, 2012. p. 47–63. (Coleção ABEPRO).

MEHTA, A. *et al*. Financial Fraud Detection through K-means Clustering-based Unsupervised Learning. **Economic and Financial Research Letters**, v. 5, n. 5, abr. 2025. Disponível em: https://www.researchgate.net/publication/394088510_Financial_Fraud_Detection_through_K-means_Clustering-based_unsupervised_Learning. Acesso em: 10 nov. 2025.

MELO, D. *et al*. Digital Government and User Experience. *In*: **SBSI - BRAZILIAN SYMPOSIUM ON INFORMATION SYSTEMS**, 18., 2022, Curitiba. New York: ACM, 2022. p. 42:1–42:9. Disponível em: https://doi.org/10.1145/3535511.3535553. Acesso em: 2 nov. 2025.

MIKOLOV, T. *et al*. **Efficient estimation of word representations in vector space**, 2013. Disponível em: https://doi.org/10.48550/arXiv.1301.3781. Acesso em: 2 nov. 2025.

MORABITO NETO, R.; PUREZA, V. Modelagem e simulação. *In*: MIGUEL, Paulo Augusto Cauchick (org.). **Metodologia de pesquisa em Engenharia de Produção e Gestão de Operações**. 2. ed. Rio de Janeiro: Elsevier, 2012. p. 169–198. (Coleção ABEPRO).

MOREIRA, J. P. F. *et al*. Licitação pública: pregão eletrônico como ferramenta de economia. **Revista Contemporânea,** v. 4, n. 4, p. 1–20, 2024. Disponível em: https://ojs.revistacontemporanea.com/ojs/index.php/home/article/view/4111. Acesso em: 2 nov. 2025.

MOTA, C. *et al*. Classificação de páginas de petições iniciais utilizando redes neurais. *In*: **ENIAC**, 13., 2020, Porto Alegre. Porto Alegre: SBC, 2020. Disponível em: https://doi.org/10.5753/eniac.2020.12139. Acesso em: 2 nov. 2025.

NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. **Journal of the American Medical Informatics Association**, v. 18, n. 5, p. 544–551, 2011. Disponível em: https://doi.org/10.1136/amiajnl-2011-000464. Acesso em: 10 nov. 2025.

NASCIMENTO, M. C. V.; CARVALHO, A. de C. P. L. F. Spectral methods for graph clustering: a survey. **European Journal of Operational Research**, v. 211, n. 2, p. 221–231, 2011. Disponível em: https://doi.org/10.1016/j.ejor.2010.08.012. Acesso em: 2 nov. 2025.

NASEEM, U. *et al*. **A Comprehensive Survey on Word Representation Models:** From Classical to State-of-the-Art Word Representation Language Models, 2020. Disponível em: https://doi.org/10.48550/arXiv.2010.15036. Acesso em: 10 nov. 2025.

NUNES, A.; DANTAS, L. de O. Eficiência do sistema de registro de preços. **Universitas: Gestão e TI**, v. 2, n. 1, p. 15–29, 2012. Disponível em: https://doi.org/10.5102/un.gti.v2i1.1627. Acesso em: 2 nov. 2025.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. **GloVe: Global vectors for word representation**. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing *(EMNLP)*. 2014.

ROUSSEEUW, P. J. Silhouettes: a graphical aid. **Journal of Computational and Applied Mathematics**, v. 20, n. 1, p. 53–65, 1987. Disponível em: https://doi.org/10.1016/0377-0427(87)90125-7. Acesso em: 2 nov. 2025.

RUSSELL, S.; NORVIG, P. **Artificial intelligence:** a modern approach. 4. ed. Hoboken, NJ: Prentice Hall, 2021.

SAMPIERI, R. H.; COLLADO, C. F.; LUCIO, M. P. B. **Metodologia de pesquisa**. 5. ed. Porto Alegre: Penso, 2013.

SANTOS, D. de A. A. *et al*. Análise da efetividade de processos de SRP. **Management Control Review**, v. 6, n. 2, p. 28–44, 2022. Disponível em: https://doi.org/10.51720/mcr.v6i2.4401. Acesso em: 2 nov. 2025.

SAKKA, F. *et al*. Human resource management in the era of AI. **Academy of Strategic Management Journal**, v. 21, Special Issue 1, p. 1–14, 2022. Acesso em: 2 nov. 2025.
SHI, J. Optimization of frozen goods distribution logistics. **Scientific Reports**, v. 14, art. 22477, 2024. Disponível em: https://doi.org/10.1038/s41598-024-72723-2. Acesso em: 2 nov. 2025.

STEINHAUS, H. Sur la division des corps matériels en parties. **Bulletin de l'Académie Polonaise des Sciences**, Classe 3, v. 4, p. 801–804, 1957. ISSN 0001-4095.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. Boston: Pearson, 2014.

THORNDIKE, R. L. Who belongs in the family? **Psychometrika**, v. 18, n. 4, p. 267-276, dez. 1953. Disponível em: https://doi.org/10.1007/BF02289263. Acesso em: 2 nov. 2025.

TOOSI, A. *et al*. A brief history of AI. **PET Clinics**, v. 16, n. 4, p. 449–469, 2021. Disponível em: https://doi.org/10.1016/j.cpet.2021.07.001. Acesso em: 2 nov. 2025.

UMARGONO, E. *et al*. K-means clustering optimization using the elbow method. *In*: **ISSTEC**, 2., 2019. p. 121–129. Disponível em: https://doi.org/10.2991/assehr.k.201010.019. Acesso em: 2 nov. 2025.

VENTORIM, I. M. *et al*. A biased sampling method for applying DBSCAN. *In*: **ENIAC**, 17., 2020, Online. Porto Alegre: SBC, 2020. Disponível em: https://doi.org/10.5753/eniac.2020.12129. Acesso em: 2 nov. 2025.

WANG, M.; HU, F. The application of NLTK library. **Theory and Practice in Language Studies**, v. 11, n. 9, p. 1041–1049, 2021. Disponível em: https://doi.org/10.17507/tpls.1109.09. Acesso em: 2 nov. 2025.

APÊNDICE A: PRODUTO TECNOLÓGICO

```
!pip install PyPDF2
!pip install nltk
!pip install gensim
!pip install numpy
!pip install scikit-learn
!pip install matplotlib
import glob
import os
import PyPDF2
import nltk
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from gensim.models import Word2Vec
import numpy as np
from sklearn.cluster import DBSCAN
from sklearn.decomposition import PCA
from sklearn.datasets import make_blobs
import matplotlib.pyplot as plt

# Função para extrair o texto de um PDF
def extract_text(pdf_file):
    pdf_reader = PyPDF2.PdfReader(pdf_file)
    text = ""
    for page_num in range(len(pdf_reader.pages)):
        page = pdf_reader.pages[page_num]
        text += page.extract_text()
    return text

# Função para pré-processar o texto
def preprocess_text(text):
    words = nltk.word_tokenize(text)
    words = [word.lower() for word in words]
    words = [word for word in words if word not in
stopwords.words('portuguese')]
    stemmer = PorterStemmer()
    words = [stemmer.stem(word) for word in words]
    return words

# Carregar os PDFs e extrair o texto
pdf_files = glob.glob(os.path.join("/content/drive/MyDrive/1.
Mestrado/Testes de Codigo/Material teste/Parcelas", "*.pdf"))
texts = [extract_text(file) for file in pdf_files]

# Remoção de stopwords
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('punkt_tab')
sentences = [preprocess_text(text) for text in texts]


# Criar um modelo Word2Vec
model = Word2Vec(sentences, vector_size=100, window=5, min_count=1,
workers=4)

# Criar representações vetoriais
def create_vector(text):
```

```
    words = preprocess_text(text)
    vectors = []
    for word in words:
        if word in model.wv:
            vectors.append(model.wv[word])
    if vectors:
        return np.mean(vectors, axis=0)
    else:
        return np.zeros(100)  # Vetor nulo se nenhuma palavra for
encontrada

vectors = [create_vector(text) for text in texts]

# Aplicar o clustering com DBSCAN
# Ajustar os parâmetros eps e min_samples conforme necessário
dbscan = DBSCAN(eps=0.5, min_samples=5).fit(vectors)
labels = dbscan.labels_

# Visualizar os clusters em um espaço 2D
pca = PCA(n_components=2)
reduced_data = pca.fit_transform(vectors)

# Criar um scatter plot colorindo os pontos de acordo com os clusters
plt.scatter(reduced_data[:, 0], reduced_data[:, 1], c=labels,
cmap='viridis')
plt.title('Clusters de Documentos')
plt.xlabel('Componente Principal 1')
plt.ylabel('Componente Principal 2')
plt.show()

from sklearn.cluster import KMeans

kmeans = KMeans(n_clusters=3, random_state=0).fit(vectors)

for i, cluster in enumerate(kmeans.labels_):
    #print(f"Cluster {i}: {pdf_files[i]}")
    #print(cluster)
    print(f"{i};{cluster};{pdf_files[i]}")

n_samples = 300
n_clusters = 3
random_state = 42
X, _ = make_blobs(n_samples=n_samples, centers=n_clusters,
cluster_std=1.0, random_state=random_state)

kmeans = KMeans(n_clusters=n_clusters, init='k-means++', max_iter=300,
random_state=random_state)
kmeans.fit(X)

novo_ponto = np.array([[3, 1]])

cluster_do_novo_ponto = kmeans.predict(novo_ponto)

labels = kmeans.labels_
centroids = kmeans.cluster_centers_

plt.figure(figsize=(8, 6))
```

```
for i in range(n_clusters):
    plt.scatter(X[labels == i, 0], X[labels == i, 1], label=f'Cluster
{i+1}')

plt.scatter(centroids[:, 0], centroids[:, 1], marker='x', s=200,
color='black', label='Centroides')

plt.scatter(novo_ponto[:, 0], novo_ponto[:, 1], marker='*', s=200,
color='red', label=f'Novo Ponto (Cluster
{cluster_do_novo_ponto[0]+1})')

plt.title('KMeans Clustering com Novo Ponto')
plt.xlabel('Componente Principal 1')
plt.ylabel('Componente Principal 2')
plt.legend()
plt.grid(True)
plt.show()

print(f"O novo ponto pertence ao cluster: {cluster_do_novo_ponto[0] +
1}")
```