TRIBUNAL DE CONTAS DO ESTADO DO CEARÁ
REVISTA Controle Doutrina e Artigos
ISSN 1980-086X e-ISSN 2525-3387
www.revistacontrole.tce.ce.gov.br

# Suspeitas de conluios em licitações no estado do Ceará: uma abordagem utilizando mineração de dados e aprendizado de máquina

Suspicions of collusion in public biddings in the state of Ceará: an approach using data mining and machine learning

https://doi.org/10.32586/rcda.v23i2.1014

**Antonio Leal Sobrinho**[1]

## RESUMO

Este artigo trata da aplicação de técnicas de mineração de dados e algoritmos de aprendizado de máquina para analisar suspeitas de conluio nas licitações públicas, nos municípios do estado do Ceará. A mineração de dados permite extrair informações relevantes de diferentes conjuntos de dados, enquanto o aprendizado de máquina possibilita a identificação de padrões complexos e anomalias que podem passar despercebidas por métodos tradicionais de auditoria. Utilizando os algoritmos de aprendizado de máquina, Apriori e *FrequentPattern-Growth* (*FP-Growth*), o estudo foca na análise de históricos de licitações, por período de gestão e relações entre empresas. O objetivo geral da pesquisa é identificar, por meio dos algoritmos Apriori e FP-Growth, suspeitas de conluios e oferecer um modelo matemático para cálculo do indicador de conluio potencial (ICP) para as empresas participantes em licitações nos municípios do estado do Ceará. Para tanto, foram utilizados dados do Sistema de Informações Municipais (SIM), do Tribunal de Contas do Ceará (TCE-CE), da Controladoria Geral da União (CGU), da Receita Federal e dos Portais da Transparência dos municípios. Indicadores de risco na participação de algumas empresas

1 Doutorando em Ciência da Computação pela Universidade Estadual do Ceará (UECE), Mestre em Economia pela Universidade Federal do Ceará (UFC), Especialista em Informática pela UECE, Especialista em Controle Externo pela Universidade Vale do Acaráu (UVA), Especialista em Auditoria Governamental pelo Instituto Plácido Castelo (IPC-TCE). Cientista de Dados, graduado pela UnicSul. Servidor do TCE-CE, lotado na Diretoria de Soluções Analíticas da Secretaria de Tecnologia da Informação. Atuou como Professor Substituto na Universidade Estadual do Ceará (Uece). Atualmente é professor na pós-graduação em Auditoria e Controle Interno (Uece) na disciplina de Auditoria de Sistemas de Informação e na graduação em Computação (Centro Universitário Farias Brito) nas disciplinas de Banco de Dados, Engenharia de Software e Lógica de Programação. E-mail: leal.sobrinho@tce.ce.gov.br

foram calculados. Além disso, os licitantes foram ranqueados pelo Indicador de Conluio Potencial (ICP), que é um indicador que ajuda a avaliar a associação entre empresas e que leva em consideração múltiplos aspectos das relações entre elas, como porte, graus de parentesco, entre outros. Os resultados apresentaram 53 empresas suspeitas, para as quais foram calculados os valores da matriz de risco das associações que essas faziam parte e os seus respectivos ICPs.

**Palavras-chave:** licitações; conluios; algoritmos; mineração de dados; aprendizado de máquina.

## ABSTRACT

This paper discusses the application of data mining techniques and machine learning algorithms to analyze suspicions of collusion in public bidding processes within municipalities in the state of Ceará. Data mining allows the extraction of relevant information from different data sets, while machine learning enables the identification of complex patterns and anomalies that may go unnoticed by traditional auditing methods. Using machine learning algorithms, Apriori and Frequent Pattern-Growth (FP-Growth), the study focuses on the analysis of bidding histories by management period, and relationships between companies. The general objective of the research is to identify, through the Apriori and FP-Growth algorithms, suspicions of collusion and to offer a mathematical model for calculating the Potential Collusion Indicator PCI for companies participating in bids in municipalities in the state of Ceará. To this end, data from the Municipal Information System (SIM), the Court of Auditors of Ceará (TCE-CE), the Union's General Comptroller (CGU), the Federal Revenue Service and Municipal Transparency Portals were used. Risk indicators in the participation of some companies were calculated. In addition, bidders were ranked by the Potential Collusion Indicator (PCI), which is an indicator that helps assess the association between companies and takes into account multiple aspects of the relationships between them, such as size, degrees of kinship, among others. The results showed 53 suspicious companies,

for which the risk matrix values of the associations to which they belonged were calculated, and their respective PCIs.

**Keywords:** bidding; collusion; algorithms; data mining; machine learning.

Avaliado pelo sistema double blind review (SEER/OJS — versão 3)

Data de submissão: 09/01/2025
Data de aprovação: 06/03/2025
Data de versão final: 29/04/2025
Data de publicação online: 23/06/2025

## 1 INTRODUÇÃO

A transparência e a integridade nas licitações públicas são pilares fundamentais para assegurar a justiça e a eficiência no uso dos recursos públicos. No entanto, a possibilidade de ocorrência de associações indevidas entre empresas proponentes tem se mostrado uma preocupação recorrente em diversas esferas governamentais. Essas associações, frequentemente tratadas como conluios, caracterizadas pela combinação entre empresas para manipular os resultados das licitações em benefício próprio, não apenas prejudicam a competitividade e a equidade do processo, mas também resultam em prejuízos significativos para os cofres públicos.

No contexto das licitações públicas, entende-se que a detecção de suspeitas e a prevenção de práticas fraudulentas são essenciais para garantir que os contratos públicos sejam adjudicados de maneira justa e eficiente. A complexidade envolvida na identificação de conluios em processos licitatórios reside no caráter dissimulado e frequentemente sofisticado dessas práticas. Métodos tradicionais de auditoria e fiscalização, embora necessários, muitas vezes se mostram insuficientes para detectar padrões ocultos e comportamentos anômalos que indicam a possível presença de fraudes. Nesse cenário, a aplicação de técnicas de mineração de dados e de algoritmos de aprendizado de máquina emergem como ferramentas promissoras. Essas tecnologias permitem a análise de grandes volumes

de dados de licitações, identificando padrões e correlações que poderiam passar despercebidos em uma análise convencional.

O uso de aprendizado de máquina para a detecção de conluios baseia-se na capacidade desses algoritmos de aprender a partir de dados históricos e identificar padrões suspeitos. Esses algoritmos reconhecem características comuns em processos licitatórios fraudados, como propostas com padrões de rodízio entre vencedores ou ofertantes de diferentes licitações, direcionamento, divisão de mercado ou consórcio entre concorrentes. Ao aplicar esses modelos aos dados de licitações municipais no Ceará, é possível identificar Regras de Associação (RAs), que são padrões formalizados de relacionamentos entre dois conjuntos de itens na forma de premissas e de resultados. As regras de associação são calculadas a partir de dados e são de natureza probabilística.

Este artigo busca explorar a combinação de técnicas de mineração de dados e aprendizado de máquina para detecção de suspeitas de conluios em licitações públicas municipais no estado do Ceará, entre os anos de 2005 e 2024, abrangendo quatro ciclos completos de gestão: 2005-2008, 2009-2012, 2013-2016, 2017-2020 e quase completamente o ciclo 2021-2024, tendo em vista que os dados coletados desse período alcançam abril de 2024. Por meio da análise de um conjunto de dados extensivo e variado, o artigo procura demonstrar como essas abordagens podem ser aplicadas para aumentar a transparência nos processos licitatórios. Além disso, o presente estudo pretende contribuir para o campo da investigação ao fornecer o Indicador de Conluio Potencial (ICP) para proponentes atuantes no Ceará que oferecem maior risco de cometimento de ilícitos. Esse indicador poderá ser utilizado em diferentes contextos e esferas governamentais, buscando mitigar a possibilidade de ações fraudulentas em licitações por parte de algumas empresas suspeitas.

## 2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção são apresentados os conceitos relacionados ao trabalho, com abordagem teórica sobre licitações públicas e detalhamento do que é aprendizado de máquina, com foco em técnicas de mineração de dados e o domínio de aplicação no contexto de processos licitatórios dos governos municipais do Ceará. Ademais, são abordados alguns trabalhos já desenvolvidos, os quais utilizam técnicas de mineração de dados e aprendizado de máquina para descoberta de suspeitas de conluios nos processos licitatórios.

### 2.1 Licitações Públicas no Brasil

No Brasil, a Lei n.º 14.133, de 1º de abril de 2021, é a legislação que atualmente regula as licitações e contratos administrativos. Ela substitui gradualmente a Lei n.º 8.666/1993, a Lei do Pregão (Lei n.º 10.520/2002) e o Regime Diferenciado de Contratações Públicas (Lei n.º 12.462/2011). A licitação pública desenvolve-se por meio de um procedimento administrativo, isto é, de uma sucessão encadeada de atos administrativos, cada qual com propósito específico e todos eles em conjunto com propósito comum (NIEBUHR, 2024).

No processo licitatório a administração pública convoca, por meio de condições estabelecidas em ato próprio (edital ou convite), empresas interessadas na apresentação de propostas para o fornecimento de bens e serviços, sendo selecionada a proposta mais vantajosa para a celebração do contrato (TCU, BRASIL, 2010).

Pelo fato de licitação ser um ato público, não pode ser tratada de maneira sigilosa. A sociedade deve ter acesso aos procedimentos referentes a uma licitação. O processo licitatório deve afastar qualquer suspeita de favorecimento e garantir que o dinheiro público seja utilizado com cautela e eficiência. A licitação é a forma mais clara de se atender aos princípios das atividades da administração pública (SOUZA, 1997).

O instituto da licitação visa escolher a alternativa que oferece maior qualidade e menores preços para a realização das atividades, garantindo que a empresa contratada cumpra com as especificações escritas e acordadas (MIRANDA, 2021).

Ainda que diferentes e variadas leis imponham certa rigidez procedimental, de modo a combater e evitar práticas ilegais em processos licitatórios, a Administração Pública ainda segue bastante susceptível a essas ocorrências. Dados publicados em relatórios de auditoria do Tribunal de Contas da União (TCU), da Controladoria Geral da União (CGU) e de diversos Tribunais de Contas estaduais, ressaltam que esquemas de práticas fraudulentas em licitações públicas são comuns e historicamente causam sérios prejuízos ao erário.

## 2.2 Fatores para a prática de conluio/cartel em licitações

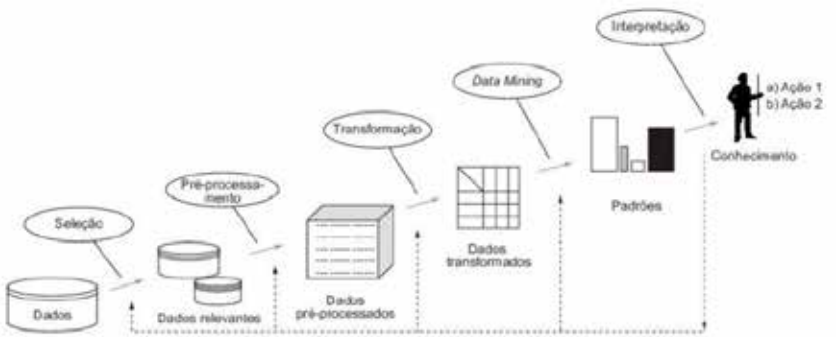
Em licitações públicas um dos grandes motivos relacionados ao desvio do dinheiro público é a formação de cartéis, que consiste em um grupo de ofertantes que fazem um acordo explícito para limitar competição entre eles em benefício próprio. O objetivo é atuar como um monopólio, mas, enquanto no monopólio a decisão cabe apenas a um ente, no cartel trata-se de uma associação voluntária de tomadores de decisão, cada qual consciente de que o seu lucro depende do comportamento de todos os ofertantes da indústria. Portanto, envolve muito mais do que uma simples definição do nível de preço e quantidade máxima do lucro (CAMPOS, 2008).

A identificação de cartel não é uma tarefa simples de se realizar, e pode incluir muitas empresas da área e os clientes raramente estão em posição de detectar a existência de um cartel, principalmente em órgãos públicos, em que o número de colaboradores destinados para esse tipo de atividade geralmente não é suficiente. Como as licitações envolvem dinheiro público, ou seja, da sociedade, é necessário auxiliar e fiscalizar a formação de cartéis por vários meios e instrumentos que se encontram disponíveis (SILVA e VIEIRA, 2023).

Braga (2015) conceitua cartéis como acordos entre concorrentes para alcançar algum tipo de benefício coletivo em detrimento da concorrência. Essas práticas geralmente visam restringir a competição, elevar preços artificialmente ou direcionar a vitória a determinados grupos ou empresas. Configuram crimes contra a administração pública e contra a ordem econômica, podendo resultar em penalidades como multas, sanções administrativas, proibição de contratar com o poder público e até prisão para os envolvidos. No Brasil, órgãos como o Conselho Administrativo de Defesa Econômica (CADE), a Controladoria-Geral da União (CGU) e os Tribunais de Contas atuam na identificação e punição desses atos.

## 2.3 Mineração de Dados

A mineração de dados é uma técnica para descobrir padrões, correlações, tendências e informações úteis a partir de grandes conjuntos de dados. É também tratada como *Knowledge-Discovery-Databases* (KDD), utilizando técnicas de estatística e aprendizado de máquina em bases de dados, a mineração de dados transforma dados brutos em conhecimento valioso.




Fonte: Fayyad et al. (1996).

Segundo *Fayyad et al.* (1996), o processo KDD refere-se ao processo de descoberta de conhecimentos úteis a partir de dados, composto

por cinco etapas, conforme Figura 1, o processo de mineração de dados está dentro do KDD e refere-se à quarta etapa desse processo. Estas etapas são distribuídas em três principais grupos: pré-processamento (seleção de dados, limpeza dos dados e tratamento de dados); processamento (mineração de dados); pós-processamento (interpretação).

Os principais objetivos da mineração de dados, além da descoberta de padrões, são: predição, classificação, agrupamento e detecção de anomalias. Podemos aplicar a mineração de dados por meio de várias estratégias para alcançar seus objetivos, entre elas: algoritmos de associação, classificação, análise de regressão, redes neurais, árvores de decisão, entre outras.

Por meio de vários campos de pesquisa, como ciência de dados, aprendizado de máquina, reconhecimento de padrões, banco de dados, estatística, e inteligência artificial, o processo de descoberta de conhecimento em bases de dados evoluiu e continua evoluindo. Consequentemente, as técnicas utilizadas em KDD não devem ser vistas como substitutas de outros métodos e áreas de análise, mas de forma que possam trabalhar juntas com o objetivo de melhorar os resultados das explorações feitas (Fayyad *et al*., 1996)

## 2.4 Aprendizado de Máquina

De acordo com Han eKamber (2005), aprendizado de máquina pode ser definido como uma área que estuda a maneira com que os computadores podem aprender ou melhorar seu desempenho de forma automática. Diferente dos sistemas de computação tradicionais, que seguem instruções programadas explicitamente, os sistemas de aprendizado de máquina usam dados para identificar padrões e construir modelos que podem ser usados para resolver problemas específicos sem a necessidade de programação explícita para cada tarefa.

O uso de aprendizado de máquina para detectar conluios em licitações pode automatizar e aprimorar a fiscalização de órgãos de controle, permitindo identificar padrões fraudulentos de forma mais eficiente. Métodos como classificação, associação, clustering, análise de grafos e séries temporais podem ajudar na identificação de empresas suspeitas e na prevenção de fraudes. Nesta pesquisa, são abordados dois algoritmos de aprendizado de máquina, o Apriori e o *FP-Growth*.

### 2.4.1 Algoritmo Apriori

O algoritmo Apriori é um dos métodos mais conhecidos para a mineração de padrões frequentes e regras de associação em grandes bases de dados. Ele foi introduzido por Agrawal *et al*. (1993), e Agrawal e Skirant (1994). Baseia-se na propriedade antimonotônica dos conjuntos de itens frequentes. Essa propriedade estabelece que, se um conjunto de itens é frequente, todos os seus subconjuntos também são frequentes. O funcionamento do algoritmo inicia-se a partir da geração de candidatos. O algoritmo começa gerando conjuntos de itens únicos (1-*itemsets*) e depois combina esses conjuntos para formar conjuntos maiores (k-*itemsets*) que podem ser candidatos a frequentes.

Após a geração dos candidatos, é realizado o cálculo de suporte. A medida de suporte, que é um hiperparâmetro informado ao modelo, é uma das métricas fundamentais usadas para avaliar a frequência de ocorrência de conjuntos de itens em um banco de dados de transações. O suporte quantifica o quão comum ou raro é um *itemset* no conjunto de dados. Especificamente, o suporte de um *itemset* é definido como a proporção de transações no banco de dados que contêm aquele *itemset*. Em cada iteração, o algoritmo conta o suporte. Após esse cálculo, ocorre a filtragem de conjuntos frequentes. Os conjuntos de itens que atendem ao suporte mínimo são considerados frequentes e são usados para gerar novos can-

didatos na próxima iteração. O processo continua até que não existam mais candidatos que possam ser gerados. O suporte também desempenha um papel crucial na definição de regras de associação (RA), que são relações frequentes identificadas entre conjunto de itens, aqui tratados como empresas participantes das licitações. Por exemplo, a regra de associação A→ B (se o item A está presente, então o item B também está presente) tem um suporte que é igual ao suporte do *itemset* {A, B}. Portanto, o suporte de uma regra de associação indica a frequência com que tanto o antecedente quanto o consequente ocorrem juntos no conjunto de dados.

Outro hiperparâmetro informado ao modelo Apriori é a confiança. Essa medida avalia a força de uma RA e mede a probabilidade de que o consequente de uma RA esteja presente em uma transação, dado que o antecedente também está presente. A confiança de uma regra A → B é a proporção de transações que contêm o item A e que contêm o item B. No contexto dessa pesquisa, transação é cada licitação contida na base de dados e o item é cada empresa participante.

A confiança é usada com o suporte para avaliar a significância das RAs geradas pelo algoritmo. Enquanto o suporte garante que as regras envolvam itens frequentes, a confiança assegura que essas regras sejam fortes e úteis na prática, indicando uma alta probabilidade de ocorrência conjunta dos itens na regra. A confiança alta indica que, na maioria das vezes em que o antecedente A ocorre, o consequente B também ocorre. O que significa afirmar que é altamente provável que a empresa B esteja presente quando A está presente. Já a confiança baixa indica que, mesmo quando o antecedente A ocorra, o consequente B raramente ocorre, o que significa que empresa B não é muito dependente da empresa A.

Outras medidas obtidas a partir da aplicação do algoritmo Apriori são o *lift* (fator de aumento) e a convicção. O *lift* mede a razão entre a probabilidade observada de itens ocorrerem juntos e a probabilidade esperada de ocorrerem juntos se forem independentes. Em outras palavras, essa medida quantifica a importância de uma RA ao comparar a ocor-

rência conjunta observada com a ocorrência conjunta esperada dos itens envolvidos. *Lift* maior que 1 indica associação positiva entre as empresas, significando que elas ocorrem juntas mais frequentemente do que seria esperado se fossem independentes. *Lift* menor que 1 indica uma associação negativa, significando que os itens ocorrem juntos menos frequentemente do que seria esperado pela sua independência. *Lift* igual a 1 indica que os itens são independentes, ou seja, a ocorrência conjunta é exatamente como seria esperado se os itens fossem independentes.

É importante acrescentar que o *lift* pode ser inflacionado por itens raros, onde a ocorrência conjunta é rara, mas, ainda assim, superior ao esperado sob independência. Por exemplo, no contexto do *dataset* desse estudo, se uma rara empresa (A) ocorre juntamente de outra empresa rara (B), em poucos casos, o *lift* pode ser muito alto, mesmo que a associação não seja significativa devido à baixa base de transações. Isso pode caracterizar um viés da medida. Outros cenários em que confiança e *lift* podem ser enviesadas são: o desequilíbrio de classes, ou empresas que participam muito raramente ou muito comumente, coocorrência acidental, ou itens que aparecem frequentemente juntos por acaso, por exemplo, empresas A, B e C que estão frequentemente juntas em transações porque são as poucas especializadas em determinado nicho de mercado.

A métrica de convicção serve para avaliar a força de uma regra de associação, comparando a frequência esperada de ocorrência de um item no conjunto de dados, quando a regra não se aplica, com a frequência observada quando a regra se aplica. A convicção mede o quão dependente é a presença de um item em relação à presença de outro item em uma regra de associação. Uma convicção igual a 1 significa que a ocorrência de B é independente da ocorrência de A. Uma convicção maior que 1 indica uma correlação positiva entre A e B, ou seja, a presença de A aumenta a probabilidade de B. Uma convicção menor que 1 indica correlação negativa entre A e B, ou que a presença de A diminui a probabilidade de B.

### 2.4.2 Algoritmo FP-Growth

O algoritmo *FP-Growth* foi concebido por Jiawei Han, Jian Pei e Yiwen Yin (2000), por meio do artigo *Mining Frequent Patterns without Candidate Generation*. No livro *Data Mining: Concepts and Techniques*, de Jiawei Han, Micheline Kamber e Jian Pei (edições de 2000, 2006 e 2011), os autores reforçam que o *FP-Growth* é uma evolução conceitual do Apriori. Outros autores como Aggarwal, C. C. (2015) na publicação *Data Mining: The Textbook*; Tan, P-N., Steinbach, M., & Kumar, V. (2005) em *Introduction to Data Mining*; e Borgelt, C. (2005) com *An Implementation of the FP-Growth Algorithm*, reforçam que o *FP-Growth* representa um avanço na mineração de padrões frequentes mostrando-se muito mais eficiente, especialmente em grandes conjuntos de dados, porque reduz drasticamente o número de combinações candidatas a serem analisadas.

A proposta do algoritmo é obter o mesmo resultado do algoritmo Apriori em termos de pesquisa dos dados, porém sem o custo computacional da etapa de geração de candidatos, vista como onerosa no algoritmo Apriori (DIZON *et. al*., 2019). Este método usa uma estratégia de dividir e conquistar para compactar os itens frequentes numa estrutura de dados chamada de *FP-tree*, divide os dados em conjuntos associados a cada item, permitindo o processamento dos itens em separado. Retorna métricas relevantes no contexto de algoritmos de associação, que são o *leverage* (alavancagem) e FB-Zhang. Essas medidas são geradas em padrões frequentes em grandes conjuntos de dados de transações.

O *leverage* é uma métrica que mede a diferença entre a frequência observada de coocorrência de dois itens e a frequência esperada se os itens fossem independentes. É uma medida útil porque identifica associações reais e ajuda a identificar se a coocorrência de itens é realmente significativa ou se é apenas o resultado do acaso. Ela complementa outras métricas e pode ser usada em conjunto com as medidas de confiança e *lift* para fornecer uma visão mais completa da força das associações. Valores

de *leverage* positivos indicam associação positiva entre os itens, ou seja, tendem a ocorrer juntos mais frequentemente do que por acaso, valores negativos indicam associação negativa, o que significa que itens tendem a ocorrer juntos menos frequentemente do que por acaso, valores zero de *leverage* indicam que os itens são independentes.

O FB-Zhang é uma medida projetada para avaliar a qualidade e a força das RAs descobertas. Essa métrica proporciona uma visão mais equilibrada e significativa da força das relações, considerando tanto a dependência positiva quanto a negativa entre os itens. Ao contrário de outras métricas que podem ser enviesadas em direção a padrões muito frequentes ou muito raros, essa métrica tenta capturar a verdadeira força da associação, levando em conta a distribuição dos itens no conjunto de dados. É útil porque leva em consideração tanto a ocorrência conjunta observada quanto a esperada, normalizando o valor máximo desses dois. Isso proporciona uma medida equilibrada que evita a supervalorização de padrões muito frequentes ou muito raros, oferecendo uma visão mais precisa da força da associação entre as empresas.

## 2.5 Trabalhos Publicados

A literatura documenta várias pesquisas em modelos baseados em mineração de dados e aprendizado de máquina para a detecção de suspeitas de conluios em licitações públicas. A Tabela 1, ilustrada a seguir, apresenta diferentes estudos baseados na aplicação de mineração e algoritmos de associação em licitações públicas.

Tabela 1 – Trabalhos publicados sobre mineração de dados e utilização de algoritmos de aprendizado de máquina para detecção de suspeitas de conluios em licitações públicas

| Título | Autor(es) | Nº | Abordagem |
|---|---|---|---|
| *AnomalyDetection: A Survey* | Varun Chandola, Arindam Banerjee, Vipin Kumar | 2009 | Apresenta uma visão geral das técnicas de detecção de anomalias, essenciais para identificar comportamentos anômalos |
| *A multi-agent data minning system for cartel detection in Brazilian government procurement* | Carlos Vinicius Sarmento Silva, Célia Ghedini Ralha | 2012 | Aborda utilização de agentes de mineração de dados com regras de associação e clusterização para detecção de cartéis em licitações |
| Aplicação do Algoritmo Apriori para Detectar Relacionamentos entre Empresas nos Processos Licitatórios do Governo Federal | Rebeca Andrade Baldomir | 2017 | Objetiva encontrar indícios de fraudes, tais como conluios e cartéis, usando algoritmo Apriori |
| *Application of machine learning in the detection of public fraud* | Marco Antonio Lopes | 2020 | Identificação de fraudes em licitações públicas com uso de aprendizado de máquina |
| Detecção de Casos Suspeitos de Conluio em Licitações Públicas: uma aplicação do algoritmo Apriori de aprendizado de máquina para o estado da Paraíba | Hilton Martins Brito Ramalho, Aléssio Tony Cavalcanti de Almeida, Alcimar Alves Fraga | 2020 | Objetiva identificar casos potencialmente suspeitos de conluio em licitações de gestões municipais da Paraíba de 2005 a 2016, utilizando o algoritmo Apriori |
| *Collusion detection in public procurement auctions with machine learning algorithms* | Manuel J. García Rodríguez, Vicente Rodríguez-Montequín, Pablo Ballesteros-Pérez, Peter E.D. Love, Regis Signor | 2022 | Esse artigo testa a precisão de onze algoritmos de aprendizado de máquina (ML) para detectar conluios usando conjuntos de dados colusivos obtidos no Brasil, Itália, Japão, Suíça e Estados Unidos |

| | | | |
|---|---|---|---|
| *A Comparative Analysis of Apriori and FP-Growth Algorithms for Market Basket Analysis Using Multi-level Association Rule Mining* | Dilara Alcan, Kubra Ozdemir, Berkay Ozkan, Ali Yigit Mucan & Tuncay Ozcan | 2023 | Nesse estudo, os algoritmos Apriori e *FP-Growth* são aplicados para análise de cesta de compras com dados reais de um varejista de FMCG |
| Descoberta de insights na análise de licitações no estado de Goiás | M.A.S. Silva, S.L. Vieira | 2023 | Relaciona possíveis indícios de irregularidades em licitações explorando *business intelligence*, *data science* e *data minning* |

Fonte: elaborada pelo autor (2024).

## 3 METODOLOGIA

Nessa seção será apresentada a metodologia utilizada para a elaboração dessa pesquisa. Inicialmente será descrito, na seção 4.1, o percurso metodológico adotado, sendo detalhados métodos e ferramentas utilizados para coleta e tratamento de dados. Na seção 4.2 serão apresentadas as formas de aplicação dos algoritmos de associação Apriori e FP-Growth para obtenção dos itens frequentes e RAs. A seção 4.3, apresenta as formulações matemáticas e os cálculos da Matriz de Riscos e do ICP. Por fim a seção de resultados seguida das considerações finais.

### 3.1 Seleção e Tratamento dos Dados

O conjunto de dados objeto da presente pesquisa foi coletado junto ao Sistema de SIM, do TCE-CE, e por meio da integração de diferentes bases de dados disponibilizados nesse órgão, como dados de empresas, pessoas físicas, funcionários de prefeituras, entre outros. Destaca-se que todo processo de coleta, leitura e integração desses dados foi realizado com o uso do Pentaho Data Integration, ferramenta criada para facilitar o processo de transferência e migração de dados de diferentes fontes e formatos. Instruções Linguagem de Consulta Estruturada (SQL), que é uma linguagem padrão utilizada para gerenciar e manipular dados em bancos

de dados relacionais, foram aplicadas em seguida, no ambiente DBeaver (ferramenta de gerenciamento de bancos de dados), como atividades de pré-processamento.

Como resultado dessas atividades, foi gerado um arquivo em formato Comma-SeparatedValues (CSV), que é um formato de arquivo usado para armazenar e transferir dados estruturados em formato tabular, como linhas e colunas, para serem utilizados, nesse contexto, pelos algoritmos de aprendizado de máquina. No intuito de não explicitar nos resultados nessa pesquisa os dados das empresas, cada empresa participante foi unicamente identificada por um número aleatoriamente gerado, com a numeração partindo de 1 até 37.183. Cada licitação também foi unicamente identificada aplicando a mesma estratégia, com a numeração iniciando em 1 até 64.982.

Foram selecionadas apenas licitações com até um item e com no máximo cinco empresas concorrendo. Essa filtragem se deu em função de que licitações com itens específicos podem atrair um maior número de pequenos fornecedores que, de outra forma, não participariam de processos licitatórios com múltiplos itens. Comumente, essas licitações são mais potencialmente dirigidas ao fornecimento de serviços mais especializados como treinamento, consultorias, entre outros.

Quanto ao número máximo de concorrentes, a razão é a utilização de valores de média nas formulações matemáticas que poderiam ser afetados por dados discrepantes de participantes por licitação. A estratégia de divisão dos dados por recortes temporais (períodos de gestão) busca obter um espaço de busca mais apropriado para a construção das RAs, tendo em vista que diferentes perfis de gestão como direcionamento político, ideologia do partido dominante, visão administrativa do gestor, podem afetar o volume de licitações, os tipos de produtos ou serviços procurados e também a formação de perfil de grupos de empresas concorrentes. A tabela abaixo resume os conjuntos de dados selecionados, segundo períodos de gestão municipal:

Tabela 2 – Dados extraídos por períodos de gestão municipal

| Período | Licitações | Participantes | Média Participantes por Licitação |
|---|---|---|---|
| 2005-2008 | 19.454 | 13.657 | 3,017 |
| 2009-2012 | 33.799 | 21.229 | 3,015 |
| 2013-2016 | 7.048 | 6.591 | 2,906 |
| 2017-2020 | 3.128 | 3.142 | 2,909 |
| 2021-2024 | 1.553 | 2.033 | 3,029 |
| Total | 64.982 | 46.552 | 2,999 |

Fonte: elaborada pelo autor (2024).

## 3.2 Aplicação dos algoritmos Apriori e FP-Growth

Os algoritmos Apriori e *FP-Growth* foram aplicados à base de dados CSV no ambiente Google Colab, por meio de códigos escritos na linguagem Python, com as bibliotecas de machine learning estendidas (*mlxtend*), e os métodos Apriori, *FP-Growth* e *Association Rules*.

A Tabela 3 apresenta métricas básicas usadas na modelagem da pesquisa:

Tabela 3 – Estatísticas básicas dos algoritmos Apriori e FP-Growth

| Nome | Estatística | Descrição |
|---|---|---|
| Suporte | $supp(X) = p(X) = \frac{n_x}{N}$ | Frequência relativa de uma empresa X no total de certames (probabilidade incondicional) onde nx é total de ocorrências para X e N é o total de transações |
| Confiança | $conf(X \rightarrow Y) = p(\frac{Y}{X}) = \frac{p(X,Y)}{p(X)}$ | Probabilidade da empresa Y participar do certame dado que a empresa X participa (probabilidade condicional) |
| *Lift* | $lift(X \rightarrow Y) = \frac{supp(X,Y)}{supp(X) \times supp(Y)}$ | Índice de desvio do suporte da RA em relação ao suporte esperado em caso de independência entre os conjuntos X e Y, onde lift> 1 indica elevado grau de associação |
| Convicção | $conv(X \rightarrow Y) = \frac{1 - supp(X,Y)}{1 - conf(X \rightarrow Y)}$ | Probabilidade de o conjunto X ser observado sem o conjunto Y, onde conv=1 implica independência entre X e Y, e conv > 1 sugere forte dependência |
| *Leverage* | $lev(A \rightarrow B) = P(A \cap B) - P(A) \times P(B)$ | Mede a diferença entre a frequência observada de coocorrência de dois itens e a frequência esperada se os itens fossem independentes |

Tabela 3 – Estatísticas básicas dos algoritmos Apriori e FP-Growth (continuação)

| Nome | Estatística | Descrição |
|---|---|---|
| FB-Zhang | $zhang(A \rightarrow B) = \frac{P(A \cap B) - P(A) \times P(B)}{max(P(A \cap B), P(A) \times P(B))}$ | Proporciona uma visão mais equilibrada e significativa da força de uma regra de associação |

Fonte: adaptada de McNicholas e Zhao (2009).

O parâmetro de suporte selecionado para aplicação dos algoritmos Apriori e FP-Growth foi de 0,2% para os anos de 2005 a 2020 e 0,3% para 2021-2024, em razão do número bastante reduzido de licitações informadas e coletadas desse último período. Isso implica dizer que para a amostra de transações 2021-2024 (1.553), uma empresa A, B ou C precisa participar ao menos de cinco licitações para constar como item frequente.

O parâmetro de confiança apontado foi de 70%, que indica a probabilidade (mínima) de participação de uma empresa C, na presença de outras empresas A e B, para que seja identificada uma regra de associação. Os dois algoritmos retornaram os mesmos itens frequentes e RAs, sendo que o *FP-Growt*h apresentou valores para as métricas de *leverage* e FB-Zhang com arredondamentos, que provocou algumas diferenças leves de valores se comparadas com resultados dessas métricas no algoritmo Apriori. Para efeito da aplicação das formulações matemáticas dessa pesquisa, foram utilizados valores resultantes da modelagem do algoritmo Apriori.

## 3.3 Cálculo da Matriz de Risco e ICP

De posse dos resultados de itens frequentes e métricas das RAs, no presente estudo, faz-se uma adaptação da função de avaliação proposta por Ralha e Silva (2012) e refinada por Ramalho et al. (2020), a fim de calcular a matriz de riscos de indicadores de conluios para RAs compreendidas no período 2021-2024, a filtragem para o período justifica-se pelo fato de que as associações presentes são entre  empresas que, muito seguramente, estão em plena atuação no atual momento, inclusive com contratos em vigência ou prestes a serem homologados.

Como contribuição da pesquisa, para todas essas empresas participantes de licitações no período anteriormente informado, individualmente, também foi calculado o ICP, que é um indicador percentual que ranqueia empresas que supostamente oferecem maior risco de cometimento de práticas inapropriadas.

A matriz de risco M(RA) possui cinco diferentes categorias, com valores de 1 a 5, indicando respectivamente os níveis de risco como Muito Baixo, Baixo, Médio, Alto e Muito Alto. A classificação segue as regras apresentadas na figura a seguir:

Figura 2 – Regras de classificação da matriz de riscos M(RA).

**Representação Lógica da Fórmula M(RA):**

1. Nível 5: Risco Muito Alto
   - Condição:
     $$(\text{PARENTESCO} \vee \text{SANCAO é verdadeiro}) \wedge (p_r > \hat{p} \vee c_r > \hat{c})$$
2. Nível 4: Risco Alto
   - Condição:
     $$(p_r > \hat{p} \wedge c_r \leq \hat{c}) \vee (\text{PARENTESCO} \vee \text{SANCAO é verdadeiro}) \wedge (p_r \leq \hat{p} \wedge c_r > \hat{c})$$
3. Nível 3: Risco Médio
   - Condição:
     $$p_r > \hat{p} \wedge c_r \leq \hat{c} \wedge (\text{PARENTESCO} \wedge \text{SANCAO são falsos})$$
4. Nível 2: Risco Baixo
   - Condição:
     $$p_r \leq \hat{p} \wedge c_r \leq \hat{c} \wedge (\text{PARENTESCO} \wedge \text{SANCAO são falsos})$$
5. Nível 1: Risco Muito Baixo
   - Condição:
     $$p_r \leq \hat{p} \wedge c_r \leq \hat{c} \wedge (\text{PARENTESCO} \wedge \text{SANCAO são falsos})$$

Fonte: elaborada pelo autor (2024).

Onde $p_r$ é a probabilidade de ao menos uma empresa da RA vencer uma licitação quando todo grupo envolvido concorre; $c_r$ é o número médio de concorrentes dos membros da RA em licitações; $p\hat{}ec\hat{}$ são parâmetros limiares (críticos) de probabilidade de o rodízio induzir vitória e de média de concorrentes; $\vee$: operador "OU" lógico. $\wedge$: operador "E" lógico.

Conforme Ralha e Silva (2012), a probabilidade $p_r = \frac{vr}{nr}$ é dada pela razão entre o número de vezes que algum fornecedor do grupo r venceu licitações em que todo o grupo participou ($vr$) e o total de licitações envolvendo participação conjunta do grupo ($nr$). Os valores de *p* e *c* seguem o proposto por Ramalho *et al*. (2020), que empregam o limiar de risco de jogo cooperativo $p = 0{,}50$ e o limiar de concorrência c igual a média histórica de proponentes por licitação para o período, nesse caso igual a 3,029.

A variável PARENTESCO indica se há relação de parentesco entre o responsável pela empresa proponente e algum funcionário da prefeitura onde se realiza o certame. Já a variável SANCAO indica se a empresa presente na RA possui alguma sanção ou impedimento de participação em licitações públicas.

O ICP de cada empresa é calculado considerando as médias de suporte, confiança, *lift*, convicção, *leverage* e zhang das RAs cujas empresas estejam presentes. Além dessas variáveis, inclui-se na fórmula a idade da empresa (tempo de atuação), a quantidade de empregados e a quantidade de atividades na Classificação Nacional de Atividades Econômicas (CNAE), que indica as especialidades que cada empresa declara possuir.

A fórmula para cálculo do ICP é mostrada a seguir:

$$\text{ICP (X)} = \text{suporte} + \text{confiança} + \text{lift} + \text{convicçao} + \text{leverage} + \text{zhang} + \text{idade} + \text{qtde\_emp} + \text{qtde\_cnae}$$

Onde:

- suporte é a média de valores de suporte para empresa X;
- confiança é a média de valores de confiança para empresa X;
- *lift* é a média de valores de *lift* para empresa X;
- convicção é a média de valores de convicção para empresa X;
- *leverage* é a média de valores de *leverage* para empresa X;

• *zhang* é a média de valores de *zhang* para empresa X;

• idade é o tempo de existência da empresa X calculado por idade = 1/Idade + 1, que indica que quanto menor o tempo de atuação da empresa, maior o risco;

• qtde_emp é quantidade de empregados declarada pela empresa X, calculado por qtde_empregados = 1/qtde_empregados + 1, quanto menos funcionários, mais risco;

• qtde_cnae é quantidade de especialidades que a empresa X afirma possuir, intuitivamente, quanto mais *cnaes*, mais risco.

## 4 RESULTADOS

Nesta seção os resultados gerados são apresentados com vistas a detectar suspeitas de conluios entre empresas nos processos licitatórios, separadamente em diferentes períodos de gestão, compreendidos entre os anos 2005 e 2024. Os resultados mostram padrões de comportamento entre empresas e indicativos de suspeição de conluio.

A Tabela 4 apresenta descritivas dos resultados da aplicação dos dois modelos de associação em todo o conjunto de dados, indicando frequência relativa, grau de associação e de independência entre empresas. Também mostra indicadores de coocorrências significativas e qualidade e força das regras de associação descobertas.

Tabela 4 – Estatísticas descritivas Itens Frequentes e RAs identificadas por períodos de gestão municipal (algoritmo Apriori)

| Período | Itens Frequentes | RAs | Médias – RAs | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | supp | Conf | lift | conv | lev | zhang |
| 2005-2008 | 310 | | 0,00255 | 0,81211 | 132,23011 | 5,03615 | 0,00251 | 0,98938 |
| 2009-2012 | 259 | | 0,00226 | 0,84235 | 176,63385 | 6,52838 | 0,00225 | 0,99638 |
| 2013-2016 | 259 | 23 | 0,00272 | 0,92354 | 299,04634 | 6,51827 | 0,00270 | 0,99847 |
| 2017-2020 | 363 | 77 | 0,00349 | 0,92167 | 226,47862 | 6,89432 | 0,00346 | 0,99675 |
| 2021-2024 | 289 | 83 | 0,00380 | 0,90051 | 148,25601 | 4,57410 | 0,00376 | 0,99381 |
| Total | 1.480 | 193 | 0,00296 | 0,88003 | 196,52898 | 5,91024 | 0,00293 | 0,99495 |

Fonte: elaborada pelo autor (2024).

Considerando todo o período, de 2005 a 2024, os algoritmos mapearam um total de 193 RAs, sendo o intervalo de 2021-2024 aquele com maior o registro de regras (83 casos, ou 43% do total). Os critérios iniciais usados para formação desses padrões para esse período basearam-se, particularmente, nas estatísticas de suporte (ao menos 0,003) e de confiança (ao menos 0,70). Para as 83 RAs do período 2021-2024, foram identificadas 53 empresas, para as quais foram calculados os valores da matriz de risco das RAs que essas faziam parte e os seus respectivos ICPs.

Os resultados apresentados da Tabela 5 contêm as empresas situadas na matriz de risco e as classificações de risco das RAs, calculados conforme valores de estatísticas apresentados na seção 3.2.

Tabela 5 – Resultados matriz de risco

| ID Empresa(s) Antecedente(s) RA | ID Empresa(s) Consequente(s) RA | Classificação de Risco |
|---|---|---|
| 20343 | 17322 | Alto |
| 21949 | 3162 | Alto |
| 6353 | 17769 | Médio |
| 8905 | 3162 | Médio |
| 9051 | 24083 | Médio |
| 11300 | 24254 | Médio |
| 11300 | 17769, 24254 | Médio |
| 15587 | 23400, 4143 | Médio |
| 13140 | 18761 | Médio |
| 14325 | 8825 | Médio |
| 14325 | 18773 | Médio |
| 14325 | 8825, 18773 | Médio |
| 15587 | 23400, 24584 | Médio |
| 16772 | 23948 | Médio |
| 16772 | 23948, 25508 | Médio |
| 16951 | 17769 | Médio |
| 18761 | 13140 | Médio |
| 18773 | 8825 | Médio |
| 18773 | 14325 | Médio |
| 18773 | 14325, 8825 | Médio |
| 19207 | 24777 | Médio |
| 19966 | 24870 | Médio |
| 21358 | 21617 | Médio |
| 21358 | 22326 | Médio |
| 21358 | 22326, 21617 | Médio |
| 21617 | 21358 | Médio |
| 21617 | 21358, 22326 | Médio |
| 22226 | 25572 | Médio |
| 22326 | 21358 | Médio |
| 22326 | 21358, 21617 | Médio |
| 22620 | 25013 | Médio |

Tabela 5 – Resultados matriz de risco (continuação)

| ID Empresa(s) Antecedente(s) RA | ID Empresa(s) Consequente(s) RA | Classificação de Risco |
|---|---|---|
| 23446 | 24083 | Médio |
| 24254 | 17769 | Médio |
| 24584 | 23400 | Médio |
| 24584 | 15587, 23400 | Médio |
| 24592 | 13981 | Médio |
| 24870 | 19966 | Médio |
| 25013 | 22620 | Médio |
| 25502 | 20739 | Médio |
| 25502 | 20739, 23948 | Médio |
| 25572 | 22226 | Médio |
| 31927 | 17769 | Médio |
| 11300, 17769 | 24254 | Médio |
| 11300, 24254 | 17769 | Médio |
| 21183, 11369 | 13981 | Médio |
| 22398, 13242 | 20739 | Médio |
| 14325, 18773 | 8825 | Médio |
| 8825, 14325 | 18773 | Médio |
| 24584 | 23400 | Médio |
| 16772, 25508 | 23948 | Médio |
| 17769, 24254 | 11300 | Médio |
| 8825, 18773 | 14325 | Médio |
| 20367, 23721 | 23400 | Médio |
| 20389, 22898 | 23948 | Médio |
| 20389, 23948 | 22898 | Médio |
| 13242, 20739 | 22398 | Médio |
| 20739, 22398 | 13242 | Médio |
| 20739, 25502 | 23948 | Médio |
| 21617, 21358 | 22326 | Médio |
| 21358, 22326 | 21617 | Médio |
| 21617, 22326 | 21358 | Médio |
| 14513, 22898 | 23948 | Médio |
| 15587, 23400 | 24584 | Médio |

Tabela 5 – Resultados matriz de risco (continuação)

| ID Empresa(s) Antecedente(s) RA | ID Empresa(s) Consequente(s) RA | Classificação de Risco |
|---|---|---|
| 20367, 23400 | 23721 | Médio |
| 23400, 23721 | 20367 | Médio |
| 23400, 24584 | 15587 | Médio |
| 23948, 14513 | 22898 | Médio |
| 16772, 23948 | 25508 | Médio |
| 20739, 23948 | 25502 | Médio |
| 23948, 25502 | 20739 | Médio |
| 23948, 25508 | 16772 | Médio |
| 11300 | 17769 | Muito Baixo |
| 11369 | 13981 | Muito Baixo |
| 15587 | 23400 | Muito Baixo |
| 15587 | 24584 | Muito Baixo |
| 16772 | 25508 | Muito Baixo |
| 18946 | 24870 | Muito Baixo |
| 21183 | 13981 | Muito Baixo |
| 21617 | 22326 | Muito Baixo |
| 22326 | 21617 | Muito Baixo |
| 22649 | 14857 | Muito Baixo |
| 24584 | 15587 | Muito Baixo |
| 25502 | 23948 | Muito Baixo |

Fonte: elaborada pelo autor (2024).

As empresas 20343, 17322, 21949 e 3162 foram classificadas com um risco alto, o que indica que possuem uma probabilidade significativa de estarem envolvidas em atividades de alto risco dentro das RAs mencionadas.

A maioria das empresas listadas encontra-se na categoria de risco médio. Alguns exemplos incluem: 6353 com 17769, 8905 com 3162, 9051 com 24083, 11300 com 24254, 15587 com 23400 e 4143. Para muitas dessas empresas, há múltiplas associações, tanto como antecedentes quanto como consequentes, o que indica interações frequentes ou ligações com outras empresas dentro das RAs. Por exemplo, a empresa 11300 aparece

em múltiplas combinações, o que pode sugerir uma posição central ou um alto grau de conectividade com outras entidades.

Empresas com risco muito baixo incluem: 11300 com 17769, 11369 com 13981, 15587 com 23400, 16772 com 25508. Essas empresas possuem uma baixa probabilidade de envolvimento em atividades de risco nas RAs. As interações entre estas empresas são provavelmente menos frequentes ou ocorrem em contextos de menor risco.

A classificação de risco apresentada é baseada em uma análise de interações entre empresas dentro das Regras de Associação. A presença de múltiplos IDs tanto como antecedentes quanto como consequentes pode indicar a complexidade das relações e a necessidade de uma análise mais detalhada para entender as razões subjacentes a essas classificações de risco.

A tabela apresentada a seguir mostra o ranqueamento das empresas segundo o Indicador de Conluio Potencial (ICP), calculado conforme fórmula anteriormente apresentada. As variáveis utilizadas são medidas estatísticas que indicam a força das associações, os níveis de dependência e correlações entre empresas. Outras variáveis utilizadas no cálculo foram a idade da empresa, a quantidade de *cnaes* cadastradas e quantidade de empregados declarados. Para facilitar a interpretação, os resultados foram normalizados, gerando valores situados entre 0 e 1.

Tabela 6 – Resultados ICP

| Posição | ID_Empresa | ICP |
|---|---|---|
| 1 | 24777 | 0.9999999709 |
| 2 | 19207 | 0.9999853983 |
| 3 | 19966 | 0.9941129937 |
| 4 | 23446 | 0.9893562362 |
| 5 | 24083 | 0.9890288569 |
| 6 | 20367 | 0.9887324665 |
| 7 | 9051 | 0.9886745375 |
| 8 | 24870 | 0.9881540242 |
| 9 | 23721 | 0.9879289082 |

Tabela 6 – Resultados ICP (continuação)

| Posição | ID_Empresa | ICP |
|---|---|---|
| 10 | 13242 | 0.9875320121 |
| 11 | 21617 | 0.9874578549 |
| 12 | 22326 | 0.9870934972 |
| 13 | 18773 | 0.9870070264 |
| 14 | 16772 | 0.9864565048 |
| 15 | 21358 | 0.986181557 |
| 16 | 23948 | 0.9860117872 |
| 17 | 25502 | 0.9853886095 |
| 18 | 22398 | 0.9851985599 |
| 19 | 24584 | 0.9851365422 |
| 20 | 20739 | 0.9850570771 |
| 21 | 25508 | 0.9848727592 |
| 22 | 23400 | 0.9841508705 |
| 23 | 25572 | 0.9830400478 |
| 24 | 22226 | 0.9830215503 |
| 25 | 15587 | 0.9827267423 |
| 26 | 20389 | 0.9807584222 |
| 27 | 25013 | 0.9789126742 |
| 28 | 22620 | 0.9788964888 |
| 29 | 22649 | 0.9761020231 |
| 30 | 18946 | 0.976094659 |
| 31 | 21949 | 0.9754013951 |
| 32 | 20343 | 0.9744670057 |
| 33 | 14513 | 0.0206217804 |
| 34 | 22898 | 0.0146428366 |
| 35 | 14325 | 0.014244362 |
| 36 | 8825 | 0.0140504387 |
| 37 | 11369 | 0.0129346766 |
| 38 | 21183 | 0.0115941314 |
| 39 | 13981 | 0.0107573955 |
| 40 | 8905 | 0.0097995394 |
| 41 | 24254 | 0.0095126452 |

Tabela 6 – Resultados ICP (continuação)

| Posição | ID_Empresa | ICP |
|---|---|---|
| 42 | 17769 | 0.0076927567 |
| 43 | 6353 | 0.0070470979 |
| 44 | 3162 | 0.0065014307 |
| 45 | 31927 | 0.0055545506 |
| 46 | 24592 | 0.0050013458 |
| 47 | 11300 | 0.004773245 |
| 48 | 14857 | 0.0040182872 |
| 49 | 18761 | 0.0036327127 |
| 50 | 17322 | 0.0022787935 |
| 51 | 16951 | 0.0021649324 |
| 52 | 13140 | 0.0019294775 |
| 53 | 4143 | 0 |

Fonte: elaborada pelo autor (2024).

Os valores de ICP variam significativamente, desde muito próximos de 1, indicando alto risco de conluio, até valores muito próximos de 0, sugerindo baixo risco. A maior parte das empresas listadas tem valores de ICP acima de 0.9, com uma queda acentuada nas últimas posições. Empresas com ICP próximo de 1 devem ser monitoradas com mais atenção devido ao alto potencial de conluio, enquanto aquelas com ICP próximo de 0 representam menor risco.

Uma análise detalhada pode ajudar na identificação de empresas que requerem maior vigilância devido ao risco de práticas de conluio, auxiliando na tomada de decisões mais informadas para mitigar esses riscos. Na Tabela 7, a seguir, são descritas estatísticas de participações e vitórias (quantitativas e percentuais) das empresas situadas no top 10 do ranqueamento do ICP.

Tabela 7 – Estatísticas de participações de empresas TOP 10 ICP

| ID EMPRESA | PARTICIPAÇÕES | VITÓRIAS | % |
|---|---|---|---|
| 19207 | 6 | 6 | 100 |
| 23721 | 12 | 12 | 100 |
| 13242 | 10 | 9 | 90 |
| 19966 | 13 | 11 | 85 |
| 24083 | 8 | 6 | 75 |
| 23446 | 5 | 2 | 40 |
| 24870 | 14 | 4 | 28 |
| 24777 | 16 | 4 | 25 |
| 20367 | 12 | 2 | 17 |
| 9051 | 5 | 0 | 0 |

Fonte: elaborada pelo autor (2024).

Os resultados apontam para possibilidades de atuação em associações ou combinações, notadamente para as empresas 19207 e 23721, que venceram todas as licitações em que participaram. Se uma empresa vence todas as licitações em que participa, pode ser necessário investigar a possibilidade de práticas anticoncorrenciais, como conluio entre concorrentes ou influências indevidas, ou pode haver uma necessidade de revisar os critérios de avaliação das propostas.

Critérios excessivamente específicos ou personalizados podem favorecer indevidamente uma única empresa. A falta de transparência pode levar a favoritismo ou manipulação dos resultados. Em mercados onde há poucas empresas qualificadas para determinados contratos, uma empresa pode vencer repetidamente devido à falta de concorrência genuína. Isso pode ocorrer em setores altamente especializados. Uma investigação faz-se necessária para avaliar cuidadosamente essas situações.

## 5 CONSIDERAÇÕES FINAIS

Essa pesquisa procurou identificar empresas suspeitas de práticas de conluio em processos licitatórios nos municípios do estado do Ceará,

entre os anos de 2005 e 2024. Por meio de mineração de dados e algoritmos de aprendizado de máquina, foi possível identificar conjuntos de empresas com suspeitas de associação indevida e classificar esses grupos conforme risco oferecido, medidos com base em dados estatísticos.

Cabe acrescentar que o indicador de conluio potencial, em conjunto com a matriz de riscos propostos nesta pesquisa, apresenta dados iniciais de riscos de potenciais conluios. O ranqueamento de empresas conforme o ICP permite um olhar crítico e análises mais cuidadosas, especialmente para aquelas situadas nas primeiras posições do ranqueamento, podendo servir como ponto de partida para maior seletividade nos processos de auditoria de empresas participantes em licitações nos municípios. É relevante acrescentar que esses resultados precisam ser avaliados com prudência, tendo em vista que essas regras de associação identificadas, a matriz de riscos, bem como o ranqueamento de ICP, precisam de análise qualitativas mais aprofundadas. Com algumas evidências obtidas nesse estudo, é possível verificar que a utilização adequada de aprendizado de máquina e mineração de dados pode contribuir substancialmente para a identificação de irregularidades e práticas inadequadas em processos licitatórios no âmbito da administração pública municipal.

Uma das limitações e ameaças à validade da pesquisa, é que a estatística de suporte (70%) não é muito alta, e o tamanho da amostra, notadamente licitações compreendidas entre 2021 e 2024 (1.553 apenas), não é significativa. De qualquer forma, os resultados indicam uma tendência e certos padrões de associações e podem ser replicados futuramente em amostras mais significativas de dados.

É importante notar que este trabalho se limitou a investigar dados de licitações públicas informadas ao Sistema de Informações Municipais (SIM), do TCE-CE. Trabalhos futuros podem aprofundar as análises buscando dados que melhor expliquem as associações entre empresas, investigar as diferenças entre licitações com discrepância entre valores estimados e valores da proposta vencedora, padrões de participação ou tempos

de submissão, das propostas para detectar comportamentos suspeitos, como submissões em série ou simultâneas. Outros estudos podem implementar uma segmentação do público a fim de investigar especificamente os ramos de negócios mais propensos a apresentar associações indevidas.

## REFERÊNCIAS

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. **Mining association rules between sets of items in large databases**. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington D.C., 1993, p. 207-216.

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules in large databases. **Proceedings of the 20th International Conference on Very Large Data Bases**, Santiago de Chile, 1994, p. 487-499.

AGGRWAL, C. **Data mining**: the textbook. Cham: Springer; 2015.

ALCAN, D.; OZDEMIR, K.; OZCAN, B.; MUCAN, A. Y.; OZCAN, T.A. A Comparative analysis of Apriori and FP-Growth algorithms for market basket analysis using multi-level association rule mining. *In*: CALISIR, F.; DURUCU, M. **Industrial Engineering in the Covid-19 Era**. GJCIE 2022. Cham: Springer, 2023. ISBN: 978-3-031-25846-6.

ARIEF, H. A; SAPTAWATI, G. A. P.; ASNAR, Y. D. W. Fraud detection based-on data mining on Indonesian E-Procurement System (SPSE). In: **2016 International Conference on Data and Software Engineering (ICoDSE)**, 2016, Denpasar, p. 1-6. DOI: 10.1109/ICODSE.2016.7936111.

BALDOMIR, R. A. **Aplicação do algoritmo Apriori para detectar relacionamentos entre empresas nos processos licitatórios do Governo Federal**. Universidade de Brasília, Brasília, 2017.

BORGEL, C. **An implementation of the FP-Growth algorithm**. Department of Knowledge Processing and Language Engineering, School of Computer Science, Otto-von-Guericke-Universityof Magdeburg, Magdeburg, 2005.

BRAGA, T. C. A. CADE, Cartéis e licitações: um novo nicho da política antitruste brasileira. **Revista de Defesa da Concorrência**, Brasília, v. 3, n. 1, p. 108-132, maio de 2015.

CAMPOS, F. As práticas de conluio nas licitações públicas à luz da teoria dos jogos. **Revista Análise Econômica**, Porto Alegre, v. 1, n. 50, p. 185-206, set, 2008.

CONTROLADORIA GERAL DA UNIÃO. **Revista da CGU**, v. 12, n. 21, jan./jun. Brasília, 2020.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: a survey. **ACM Computing Surveys (CSUR)**, v. 41, 2009. DOI: https;// doi.org/10.1145/1541880.1541882.

DIZON, F. S. V. *et al*. Learning of high dengue incidence with clustering and FP-Growth algorithm using WHO historical data. **Computing Research Repository (CoRR)**, v. 1901, n. 11376. 2019.

FAYYAD, U.; SHAPIRO, G. P.; SMYTH P. From data mining to knowledge discovery in databases. **AI Magazine**, [S. l.], v. 17, n. 3, p. 37, 1996. DOI: 10.1609/aimag.v17i3.1230. Disponível em: https://ojs.aaai.

org/aimagazine/index.php/aimagazine/article/view/1230. Acesso em: 25 jun. 2024 Acesso em: 17 mar. 2025.

HAN, J.; PEI, J.; YIN, Y. Mining frequent patterns without candidate generation. In: **Proceedings of 2000 ACM SIGMOD International Conference on Management of Data**. New York: Association for Computing Machinery, 2000, p. 1-12. DOI: 10.1145/342009.335372.

LOPES, M. A.; MONTINI, A. de Á.; COSTA, L. dos S. Application of machine learning in the detection of public fraud. In: **Anais**. São Paulo: TECSI/EAC/FEA/USP, 2020. Disponível em: https://doi.org/10.5748/17CONTECSI/PSE-6642. Acesso em: 25 jun. 2024.

MCNICHOLAS, D.; ZHAO, Y. **Association rules**: an overview. London: IGI Global, 2009.

MIRANDA, H. S. **Licitações e contratos administrativos**. Revista dos Tribunais, 5a ed. São Paulo, 2021.

NIEBUHR, J.M. Licitação pública e contrato administrativo. Ed. 7. Belo Horizonte: Fórum, 2024.

RALHA, C. G.; SARMENTO SILVA, C. V. A multi-agent data minning system for cartel detection in Brazilian Government Procurement. **Expert Systems with Application**, v. 39, n. 14, p. 11642-11656, 2012.

RAMALHO, H.M.B.; ALMEIDA, A.T.C.; FRAGA, A.A. Detecção de casos suspeitos de conluios em licitações públicas: uma aplicação do algoritmo a Priori de aprendizado de máquina para o Estado da Paraíba. In: **Teoria e Prática em Administração**, v. 10, n. 2, p. 5-22, 2020. DOI:

10.21714/2238-104X2020v10i2-51526. Disponível em: https://periodicos.ufpb.br/index.php/tpa/article/view/51526. Acesso em: 25 jun. 2024.

RODRÍGUEZ, M. J. G.; RODRÍGUEZ-MONTEQUÍN, V.; BALLESTEROS-PÉRES, P.; LOVE, P. E. D.; SIGNOR, R. Collusion detection in public procurement 100 auctions with machine learning algorithms. In: **Automation in Construction**, v. 133, p. 104047, 2022.

SILVA, M. A. S.; VIEIRA, S. L. Descoberta de insights na análise de licitações no estado de Goiás. In: **Revista Brasileira de Criminalística**, v. 12, n. 5, p. 25-38, 2023. DOI: 10.15260/rbc.v12i5.600. Disponível em: https://revista.rbc.org.br/index.php/rbc/article/view/600. Acesso em: 12 jun. 2024.

SOUZA, F. R. de. **Manual básico de licitação**. São Paulo: Nobel, 1997.

TAN, P-N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. Boston: Addison-Wesley, 2005.

TRIBUNAL DE CONTAS DA UNIÃO. Licitações e contratos, orientações e jurisprudência do TCU. 4. ed. Brasília: **TCU**: Revista ampliada e atualizada, 2010, 901p.