

Aprimoramento de modelos para detecção de conluios em licitações públicas brasileiras com variáveis estatísticas e modelos explicáveis

Lucas D. Scoralick¹, Diego N. Brandão¹, Kele T. Belloze¹

¹Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ)
Rio de Janeiro – RJ – Brazil

lucas.scoralick@aluno.cefet-rj.br
{diego.brandao, kele.belloze}@cefet-rj.br

Abstract. Collusions are secret agreements or combinations between two or more parties, usually to deceive or harm third parties. The practice of collusion in public tenders disrupts the market price balance, negatively impacting the costs and quality of public services. This study proposes a methodology to improve collusion classification models, using statistical variables combined with the analysis of explainable models to interpret results better. The results showed a significant prediction improvement of 1 to 4%, except for the decision tree algorithm.

Resumo. Conluios são acordos ou combinações secretas entre duas ou mais partes, geralmente para enganar ou prejudicar terceiros. A prática de conluios em licitações públicas perturba o equilíbrio de preços do mercado, impactando negativamente tanto os custos quanto a qualidade dos serviços públicos. Neste estudo, propõe-se uma metodologia para aprimorar os modelos de classificação de conluio, utilizando variáveis estatísticas combinadas com a análise de modelos explicáveis para melhor interpretação dos resultados. Os resultados mostraram uma sensível melhora de 1 a 4% na predição, exceto para o algoritmo de árvore de decisão.

1. Introdução

Conluios são combinações entre partes em busca de vantagens, muitas vezes relacionadas a atuação de cartéis. Tipicamente as vantagens estão associadas a preços inexequíveis, superfaturados ou sobrepreços. A existência de conluios tendem a transformar o equilíbrio de preços do mercado. Essas práticas ferem o objetivo da Lei de licitação e contratação para as Administrações Públicas (14.133/21)¹. O artigo 11 da referida lei cita como objetivos do processo licitatório assegurar a seleção da proposta mais vantajosa, isonomia entre licitantes, justa competição, evitar preços inexequíveis, superfaturados ou sobrepreços.

Dados os potenciais prejuízos à ordem econômica provocados por conluios, sua identificação é fundamental tanto para investigações de órgãos de controle quanto para empresas do setor público e privado. Essas entidades podem adotar estratégias de veto nos processos de contratação de participantes envolvidos nesse tipo de fraude. No entanto, identificar conluios não é uma tarefa fácil. Uma abordagem possível é a utilização de

¹Lei 14.133/21: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/14133.htm

algoritmos de classificação, que permitem o aprendizado de modelos a partir de atributos dos processos licitatórios.

Considerando que conluíus alteram o equilíbrio de preços do mercado, análises do comportamento estatístico podem auxiliar na identificação dessas perturbações. Este estudo explora essa característica, propondo a adição de variáveis estatísticas ao conjunto de dados de licitações, a fim de aprimorar os modelos de aprendizado. Adicionalmente, utiliza-se o método de *SHapley Additive exPlanations* (SHAP) para demonstrar a contribuição das variáveis estatísticas para o modelo. A aplicação desta metodologia resultou em modelos de aprendizagem com acurácia de 1% a 4% superior com a introdução das variáveis estatísticas. A contribuição das variáveis estatísticas foi confirmada com a aplicação do método de SHAP. As demais seções do trabalho englobam Trabalhos Relacionados, Metodologia, Resultados e Considerações Finais.

2. Trabalhos relacionados

A literatura cita diversas abordagens para detecção de conluíus. Uma das abordagens é a análise econométrica, onde o comportamento das licitações de um grupo de empresas é comparado com um grupo de controle [Porter and Zona, 1999; Busu and Busu, 2021]. Outra é a análise estatística, como executada por Abrantes-Metz et al. [2006], que analisaram o movimento de preços e indicadores estatísticos que podem caracterizar existência de conluio. Jiménez and Perdiguero [2012] analisaram e confirmaram, estatisticamente, que há uma relação positiva entre monopólio e rigidez de preços. Silva et al. [2022] propôs um *framework* que avalia se as licitações estão fora dos limites legais.

Há artigos relacionados que combinam variáveis estatísticas e aprendizado de máquina para investigar conluíus, como proposto por este trabalho. Alguns trabalhos utilizam dados rotulados [Silveira et al., 2022; Wallimann et al., 2023; Huber and Imhof, 2019; García Rodríguez et al., 2022]. Em contrapartida, aqueles trabalhos que não usam dados rotulados, normalmente aplicam métodos de agrupamentos para separar os conjuntos [Silveira et al., 2023; Júnior et al., 2023; Imhof and Wallimann, 2021].

Todos os artigos relacionados utilizam Coeficiente de Variação (CV) e Assimetria (SKEW) nos experimentos. Silveira et al. [2022] afirmam ter encontrado evidências que SKEW, Curtose (KURT) e Dispersão (SPREAD) aumentam o desempenho dos algoritmos. Huber and Imhof [2019] concluíram que as variáveis mais valiosas são CV e distância normalizada, enquanto que Silveira et al. [2023] avalia que CV e SKEW são as características consistentemente observadas nos cartéis de fraudes e licitação.

A utilização de modelos explicáveis, tais como SHAP, *Local Interpretable Model-Agnostic Explanation* (LIME) e *moDel Agnostic Language for Exploration and eXplanation* (DALEX), são referenciados em alguns artigos, normalmente associados à área de saúde [Wang et al., 2023; Nguyen and Byeon, 2023; Du et al., 2023]. No entanto, não há na literatura aplicação de modelos explicáveis em detecção de conluio.

3. Metodologia

A metodologia proposta para o desenvolvimento deste trabalho segue as etapas ilustradas na Figura 1, as quais são explicadas na sequência.

Coleta de dados. O Conselho Administrativo de Defesa Econômico (CADE) tem como uma de suas funções o combate a cartéis. Suas responsabilidades, determinadas pela

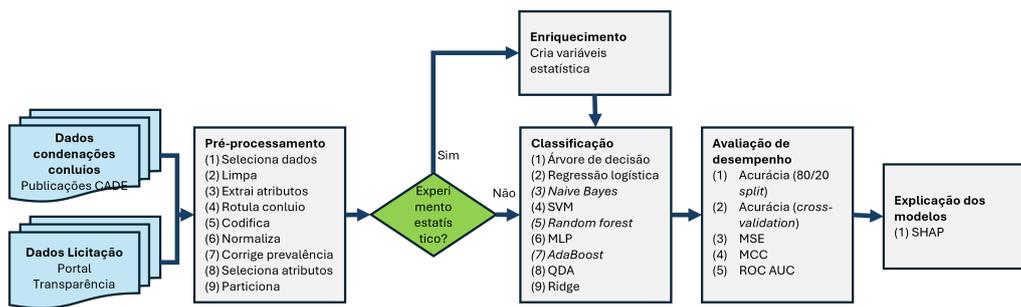


Figura 1. Metodologia para aprimoramento da identificação de conluio

Lei 12529/11², são de julgar e punir administrativamente infrações à ordem econômica, analisar e evitar excessiva concentração e promover a competitividade. Neste trabalho foram coletados dados de condenações de conluios publicados pelo CADE, vigentes no período de 2013 a 2018, conforme Tabela 1. Os dados de licitação são referentes à Administração Pública Federal e foram obtidos no Portal da Transparência do Governo Federal do Brasil³.

Tabela 1. Levantamento dos processos de condenação do CADE a partir de 2013.

Processo administrativo (CADE)	Período	Assunto
08012.010022/2008-16	2008-2013	Merenda escolar
08700.003390/2016-60	2006-2013	Tubos e conexões de PVC
08700.003396/2016-37	2004-2015	Tubos e conexões
08700.004095/2020-15	2014-2018	Produtos e serviços de inteligência de negócios (<i>business intelligence</i>)
08700.004248/2019-82	2011-2014	
08700.004455/2016-94	1999-2014	Material escolar
08700.004617/2013-41	1999-2013	Licitação de trens e metrô
08700.005637/2020-69	2017	Fornecimento de combustível
08700.005789/2015-02	2008-2014	Aquisição de sacos de lixo
08700.007278/2015-17	05/2014-11/2014	Licitações da infraero

Pré-processamento. As atividades envolveram a integração de dados dispersos em planilhas diversas⁴, exclusão de atributos com muitos valores únicos, pouca variabilidade ($<0,001$) e semanticamente similares. As licitações sem concorrentes e com quantidade de itens igual a zero foram excluídas. O atributo *valor do item* foi transformado para *valor unitário do item* de modo a explicitar a semântica do dado. Foram rotulados 1.077 registros como “Sim” para conluio, sendo estas as licitações com mais de um participante condenado para períodos coincidentes; e rotulados como “Não” 33.848.529 registros. A codificação de atributos categóricos utilizou *LabelEncoder* e *MinMaxScaler* para normalização. Duas estratégias de particionamento de dados para validação do modelo foram utilizadas, a primeira dividiu o conjunto de dados entre treinamento e teste na proporção de 80/20(%), e, a segunda utilizou o *k-fold Cross-Validation*, com 10 *folds*. Para correção de prevalência, utilizou-se a técnica *Smote* e para redução de dimensionalidade (seleção de atributos), aplicou-se χ^2 (qui-quadrado).

Enriquecimento (variáveis estatísticas). Foram criadas variáveis estatísticas a

²Lei 12.529/11: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/112529.htm

³Dados disponibilizados pela Controladoria Geral da União (CGU) no endereço eletrônico: <https://portaldatransparencia.gov.br/download-de-dados/licitacoes>

⁴Os dados de licitação do Portal Transparência são divididos em planilhas mensais de Licitação, Item de Licitação e Participantes

partir do *valor unitário dos itens*: Coeficiente de variação (CV), mede a dispersão resultante da relação do desvio padrão e média; Curtose (KURT), indica até que ponto a curva de frequência de distribuição é mais achatada que a curva normal; Assimetria (SKEW), mede o grau de deformação de uma curva de frequência de distribuição; *Outlier* (OUT), classifica se é um dado que se diferencia dos demais. Ainda, o atributo Quantidade de concorrentes (QTDC) foi adicionado.

Classificação. Os experimentos foram executados em nove diferentes algoritmos: Árvore de Decisão (DT), Regressão Logística (RL), *Naive Bayes* (NB), *Support Vector Machine* (SVM), *Random forest* (RF), *Multilayer Perceptron* (MLP), *AdaBoost*, *Quadratic Discriminant Analysis* (QDA), e Ridge (RD).

Avaliação de desempenho. Foram geradas as métricas: Acurácia, Precisão, *Recall*, F1, *Mean Squared Error* (MSE), *Matthews correlation coefficient* (MCC) e *Area under the ROC Curve* (ROC AUC). Também foram medidas a Acurácia e *Recall* na aplicação do método de validação cruzada com 10 *folds*.

Explicação dos modelos. Para os modelos, acrescidos de variáveis estatísticas, foi executado o método *SHapley Additive exPlanations* (SHAP), que ordena e apresenta quais atributos mais contribuíram para a classificação.

4. Resultados

No pré-processamento foram avaliadas as melhores abordagens para correção de prevalência (*Under Sampling*, *Over Sampling* e *Smote*) e redução de dimensionalidade. Para correção de prevalência, o *Smote* apresentou resultados mais balanceados para os experimentos com e sem variáveis estatísticas. Nesta avaliação foi utilizado o classificador SVM e as medidas acurácia média, precisão classificador negativo e precisão classificador positivo, obtendo como resultados com as variáveis estatísticas: *Under Sampling* (0,835 / 0,81 / 0,86); *Over Sampling* (0,820 / 0,76 / 0,87); *Smote* (0,835 / 0,77 / 0,87). No experimento sem variáveis estatísticas: *Under Sampling* (0,782 / 0,79 / 0,77), *Over Sampling* (0,796 / 0,74 / 0,86), *Smote* (0,795 / 0,74 / 0,85).

Para avaliação do método de seleção de atributos, foi utilizado o classificador SVM e a acurácia média como medida de avaliação. Os algoritmos de Seleção Univariada (SU), χ^2 e Análise de Componentes Principais (PCA), tiveram acurácia de 0,83 e o Eliminação Recursiva de Atributos (RFE) de 0,82. No entanto, o χ^2 obteve um *recall* médio de 0,84, enquanto que o SU, RFE e PCA foi 0,83.

Os experimentos foram executados em uma máquina *Windows 11 Home*, processador *11th Gen Intel(R) Core(TM) i7-1165G7*, 16 GB de RAM, em um ambiente *jupyter notebook 6.4.8*, *Python 3.9.12*, com as bibliotecas *pandas 1.4.4*, *numpy 1.21.5*, *matplotlib 3.5.1*, *scikit-learn 1.0.2* e *shap 0.42.1* (referências indicadas na nota de rodapé⁵).

O conjunto de dados treinado possui 33.849.606 registros e cinco atributos (Modalidade Compra, Unidade Federação, Unidade Gestora, Resultado Compra Ano, Nome Participante). No experimento com variáveis estatísticas são acrescidos mais cinco atributos (CV, KURT, SKEW, OUT e QTDC). Sobre uso de variáveis estatísticas mencionados na seção 2, em linha com os trabalhos relacionados, são utilizadas as características

⁵<https://jupyter.org/>, <https://pandas.pydata.org/>, <https://numpy.org/>, <https://matplotlib.org/>, <https://scikit-learn.org/stable/>, <https://shap.readthedocs.io/en/latest/>

estatísticas CV, KURT e SKEW. A característica OUT foi incluída para testar sua contribuição, uma vez que ela não foi utilizada em nenhum outro trabalho.

Os seguintes algoritmos tiveram acurácia média inferior a 90% para os experimentos com e sem variáveis estatísticas: QDA (0,84%/0,81%), RD (0,82%/0,78%), SVM (0,82%/0,78%) e RL (0,80%/0,78%). Os resultados dos algoritmos com melhor desempenho podem ser visualizados na Tabela 2. Todos os algoritmos tiveram desempenho superior entre 1% a 4% de aumento na acurácia com a inclusão das variáveis estatísticas, exceto DT. Da mesma forma, as métricas ROC AUC, MCC e MSE são superiores para execução com variáveis estatísticas. Curiosamente, a métrica ROC AUC para DT é superior com a inclusão das variáveis estatísticas, demonstrando que mesmo possuindo acurácia média inferior, a probabilidade de classificar corretamente é maior com a inclusão das variáveis estatísticas.

Tabela 2. Comparação dos resultados com e sem variáveis estatísticas na validação 80% treinamento e 20% teste. A acurácia dos modelos com as variáveis estatísticas é superior em todos os casos, exceto DT.

Métrica	COM variáveis estatísticas				SEM variáveis estatísticas			
	DT	RF	AdaBoost	MLP	DT	RF	AdaBoost	MLP
precisão	0,89	0,93	0,95	1,00	0,91	0,92	0,93	0,98
recall	0,89	0,93	0,95	1,00	0,90	0,92	0,93	0,98
<i>f1-score</i>	0,89	0,93	0,95	1,00	0,90	0,92	0,93	0,98
acurácia	0,89	0,93	0,95	1,00	0,90	0,92	0,93	0,98
ROC AUC	0,934	0,978	0,987	1,000	0,928	0,975	0,981	0,998
MSE	0,112	0,072	0,054	0,004	0,104	0,080	0,067	0,023
MCC	0,779	0,856	0,890	0,992	0,804	0,839	0,865	0,953

Ao utilizar o método de validação *10-Fold-Cross-Validation*, a acurácia média dos algoritmos é reduzida. Os resultados dos algoritmos com melhor desempenho podem ser visualizados na Tabela 3. Os demais tiveram acurácia para os experimentos com e sem variáveis estatísticas, respectivamente de: QDA (0,81%/0,76%), RL (0,74%/0,70%), SVM (0,74%/0,68%) e RD (0,74%/0,67%). No entanto, a diferença percentual entre a execução com e variável estatística é ainda maior, variando de 2% a 7%.

Tabela 3. Comparação dos resultados com e sem variáveis estatísticas na validação por 10-Fold-Cross-Validation. A acurácia dos modelos com as variáveis estatística é superior em todos os casos, exceto DT.

Métrica	COM variáveis estatísticas				SEM variáveis estatísticas			
	DT	RF	AdaBoost	MLP	DT	RF	AdaBoost	MLP
acurácia	0,658	0,837	0,770	0,916	0,716	0,784	0,750	0,850
<i>recall</i>	0,933	0,904	0,968	0,998	0,936	0,920	0,957	0,987

Para avaliar a contribuição das variáveis estatísticas para o modelo de predição, foi executado o método SHAP para os seguintes resultados: algoritmo de pior desempenho (DT) e algoritmo de melhor desempenho geral (MLP), ambos com variáveis estatísticas em relação ao experimento sem essas variáveis. Em relação ao DT, os quatro atributos com maior contribuição são UF, CV, QTDC e SKEW. Para o MLP são KURT, SKEW, UF e Resultado Compra Ano, como apresentado na Figura 2.

O gráfico ilustrado na Figura 2 apresenta um resumo, onde, para cada atributo as observações do conjunto de dados são representadas como pontos. Maiores concentrações de pontos são demonstradas com estes pontos crescendo na vertical. O eixo vertical, lista as variáveis do modelo por ordem de importância. O eixo horizontal são os valores resultantes do SHAP. Valores positivos indicam suporte para atributo alvo (é conluio), e o negativo para o oposto (não é conluio). As cores, representam o valor da variável, quanto mais vermelho, maior o valor. Ter essas cores em lados opostos indica que a variável é uma boa preditora, pois só em mudar o valor o modelo já consegue verificar a contribuição para a classe. O intervalo de alcance também é importante (mais a direita, maior a contribuição para o resultado) e variáveis com maior dispersão no eixo, tendem a ser mais importantes. A capacidade do SHAP de indicar a contribuição dos atributos para o modelo, ajuda na explicação e caracterização dos conluios.

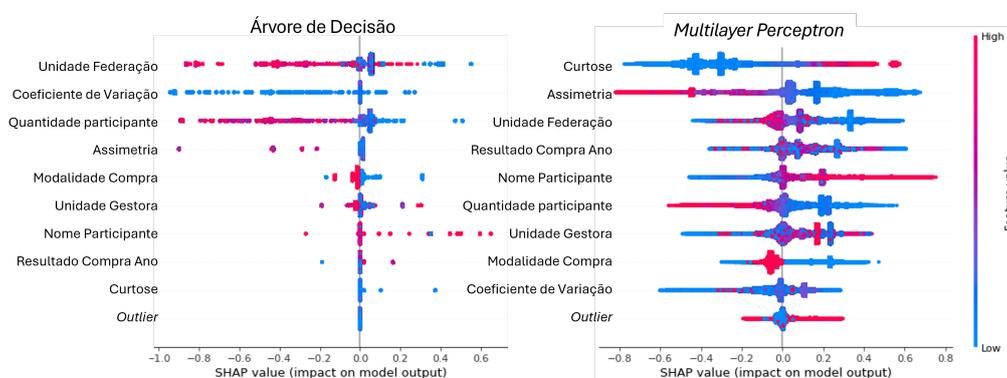


Figura 2. Gráfico de resumo do método SHAP para os algoritmos DT e MLP.

5. Considerações finais

A adição de variáveis estatísticas contribuiu positivamente para a melhoria na predição de 1% a 4% de acurácia em todos os modelos experimentados, com exceção da Árvore de Decisão. O método SHAP permitiu apresentar a contribuição das variáveis estatísticas para o modelo, confirmando a relevância da Curtose e Assimetria. A adição do atributo *Outlier* não provou ter relevância no melhoramento das predições. Estes resultados comprovam que o comportamento estatístico da ação de grupos em conluio são similares, independente do nicho que ele opera e podem ser melhor descritos a partir de componentes estatísticos em relação ao preço contratado dos produtos e serviços.

Este artigo contribui com a produção de um conjunto de dados rotulados de licitações Públicas Federais do ano de 2013 a 2018⁶, com diversos produtos e serviços. O modelo gerado pode ser usado na predição de conluios dos dados mais recentes disponibilizados pela CGU, que mantém os mesmos atributos. Para outros conjuntos de dados, sugere-se a aplicação de metodologia para treinar o modelo com o conjunto de atributos mais relevantes para o cenário. Em trabalhos futuros, sugere-se testar a contribuição de outras variáveis estatísticas e das especializações do conjunto de dado. O conjunto de dados utilizado não considerou possíveis diferenças entre os mesmos produtos, tais como marcas e tamanho da embalagem, por exemplo, o que em alguns casos pode ter grande variação.

⁶Repositório: <https://github.com/lucasdaflon/conluio-licitacoes/tree/main>

Referências

- Abrantes-Metz, R. M., Froeb, L. M., Geweke, J., and Taylor, C. T. (2006). A variance screen for collusion. *International Journal of Industrial Organization*, 24(3):467 – 486.
- Busu, M. and Busu, C. (2021). Detecting bid-rigging in public procurement. a cluster analysis approach. *Administrative Sciences*, 11(1).
- Du, Y. et al. (2023). Towards an explainable clinical decision support system for large-for-gestational-age births. *PLoS ONE*, 18(2 February).
- García Rodríguez, M. J. et al. (2022). Collusion detection in public procurement auctions with machine learning algorithms. *Automation in Construction*, 133:104047.
- Huber, M. and Imhof, D. (2019). Machine learning with screens for detecting bid-rigging cartels. *International Journal of Industrial Organization*, 65:277 – 301.
- Imhof, D. and Wallimann, H. (2021). Detecting bid-rigging coalitions in different countries and auction formats. *International Review of Law and Economics*, 68.
- Jiménez, J. L. and Perdiguero, J. (2012). Does rigidity of prices hide collusion? *Review of Industrial Organization*, 41(3):223 – 248.
- Júnior, D. P. G., de Sousa Filho, G. F., and dos Anjos Formiga Cabral, L. (2023). Classificação de fraudes em licitações públicas através do agrupamento de empresas em conluíus. *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico*.
- Nguyen, H. V. and Byeon, H. (2023). Prediction of parkinson’s disease depression using lime-based stacking ensemble model. *Mathematics*, 11(3).
- Porter, R. H. and Zona, J. D. (1999). Ohio school milk markets: An analysis of bidding. *RAND Journal of Economics*, 30(2):263 – 288.
- Silva, E., Fragoso, G., Rodrigues, L., Freitas, N., and Vinuto, T. (2022). Exposer: Framework para detecção de anomalias em licitações públicas. In *Anais Estendidos do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 77–82, Porto Alegre, RS, Brasil. SBC.
- Silveira, D., de Moraes, L. B., Fiuza, E. P., and Cajueiro, D. O. (2023). Who are you? cartel detection using unlabeled data. *International Journal of Industrial Organization*, 88.
- Silveira, D., Vasconcelos, S., Resende, M., and Cajueiro, D. O. (2022). Won’t get fooled again: A supervised machine learning approach for screening gasoline cartels. *Energy Economics*, 105.
- Wallimann, H., Imhof, D., and Huber, M. (2023). A machine learning approach for flagging incomplete bid-rigging cartels. *Computational Economics*, 62(4):1669 – 1720.
- Wang, Z. et al. (2023). Developing an explainable machine learning model to predict the mechanical ventilation duration of patients with ards in intensive care units. *Heart and Lung*, 58:74 – 81.