UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
CIÊNCIA DA COMPUTAÇÃO

Leonardo Vieira Nunes

Detecção de conluio em licitações utilizando algoritmos de machine learning

Florianópolis
7 de julho de 2024

Leonardo Vieira Nunes

Detecção de conluio em licitações utilizando algoritmos de machine learning

Trabalho de Conclusão de Curso submetido ao Curso de Graduação em Ciência da Computação do Centro Tecnológico da Universidade Federal de Santa Catarina como requisito para obtenção do título de Bacharel em Ciência da Computação.
Orientador: Prof. Jônata Tyska Carvalho, Dr.

Florianópolis
7 de julho de 2024

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Nunes, Leonardo Vieira
Detecção de conluio em licitações utilizando algoritmos de machine learning / Leonardo Vieira Nunes ; orientador, Jônata Tyska Carvalho, 2024.
67 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Ciências da Computação, Florianópolis, 2024.

Inclui referências.

1. Ciências da Computação. 2. Machine Learning. 3. Licitações Públicas. 4. Conluio. I. Carvalho, Jônata Tyska. II. Universidade Federal de Santa Catarina. Graduação em Ciências da Computação. III. Título.

Leonardo Vieira Nunes

Detecção de conluio em licitações utilizando algoritmos de machine learning

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de Bacharel em Ciência da Computação e aprovado em sua forma final pelo curso de Graduação em Ciência da Computação.

Florianópolis, 7 de julho de 2024.

Profª. Lúcia Helena Martins Pacheco, Dra.
Coordenadora do Curso

Banca Examinadora:

Prof. Jônata Tyska Carvalho, Dr.
Orientador
Universidade Federal de Santa Catarina

Profª. Jerusa Marchi, Dra.
Avaliador
Universidade Federal de Santa Catarina

Prof. Márcio Bastos Castro, Dr.
Avaliador
Universidade Federal de Santa Catarina

Dedico este trabalho à minha querida mãe, Salete; ao meu querido pai, Benedito; e à minha querida companheira, Carol; e a todos os demais membros de minha família.

AGRADECIMENTOS

Gostaria de agradecer a Deus por ter me amado desde toda a Eternidade e por ter continuamente me dado forças para seguir adiante. Também não poderia deixar de agradecer em especial à minha mãe, que me deu todo o suporte emocional para que pudesse fazer essa graduação. Espero que este trabalho possa ter honrado todo amor de mãe que tens comigo. Além disso, agradeço à Carol, a qual a vida me presenteou como a melhor amiga e companheira de vida, obrigado por estar sempre comigo, na alegria e na tristeza. Agradeço também ao professor Jônata Tyska e todos os membros do projeto CÉOS que me ajudaram tirando dúvidas para que este trabalho pudesse ser implementado. Por fim, agradeço aos colegas que conheci, dentro e fora do curso, que me proporcionaram boas experiências e companheirismo.

RESUMO

O conluio consiste em uma prática criminosa onde empresas concordam secretamente valores de lances entre si em determinada licitação com o objetivo de maximizarem o seu lucro. Esse tipo de prática inflaciona os preços e prejudica a qualidade dos produtos e serviços adquiridos pela administração pública, impactando diretamente na vida dos cidadãos. Nesse contexto, surgiram abordagens utilizando métodos baseados em aprendizado de máquina que analisam grandes quantidades de dados para detectarem a presença de conluio. Todavia, a falta de um fluxo pré-estabelecido para analisar e comparar o comportamento dos algoritmos e conjuntos de dados torna a tomada de decisão mais desafiadora. Diante disso, este estudo busca implementar um fluxo de trabalho para auxiliar na tomada de decisão em relação aos melhores algoritmos de *machine learning* a serem utilizados para determinado conjunto de dados. Para isso, foram reproduzidos resultados encontrados na literatura, otimizados os hiperparâmetros dos modelos, selecionado o que obteve a melhor acurácia balanceada e, por fim, analisado a importância das features, distribuição das classificações pelas *features* e o teste U de *Mann-Whitney*. A partir desses resultados, concluiu-se que há a necessidade de enriquecimento em conjuntos menores, uma vez que resultou em um alto desvio padrão da acurácia balanceada. Também notou-se que houve um ganho de desempenho significativo nos modelos treinados com conjuntos maiores, apresentando uma acurácia balanceada em torno de 83% para o de St. Gallen e Graubünden.

Palavras-chave: Machine Learning. Licitações Públicas. Conluio

ABSTRACT

The collusion consists of a criminal practice where companies secretly agree on bid values among themselves in a particular tender with the aim of maximizing their profit. This type of practice inflates prices and harms the quality of products and services acquired by the public administration, directly impacting the lives of citizens. In this context, approaches using machine learning methods have emerged that analyze large amounts of data to detect the presence of collusion. However, the lack of a pre-established workflow to analyze and compare the behavior of algorithms and datasets makes decision-making more challenging. Therefore, this study aims to implement a workflow to assist in decision-making regarding the best machine learning algorithms to be used for a specific dataset. To achieve this, results found in the literature were reproduced, the hyperparameters of the models were optimized, the one with the best balanced accuracy was selected, and finally, the importance of the features, distribution of classifications by features, and Mann-Whitney U Test were analyzed. From these results, it was concluded that there is a need for enrichment in smaller datasets, as it resulted in a high standard deviation of balanced accuracy. It was also noted that there was a significant performance gain in models trained with larger datasets, with a balanced accuracy of around 83% for St. Gallen and Graubünden.

Keywords: Machine Learning. Public Procurement. Collusion

LISTA DE FIGURAS

Figura 1 – Fluxograma da metodologia realizada neste trabalho 30
Figura 2 – Boxplot da reprodução dos resultados originais 33
Figura 3 – Barra de erro da reprodução dos resultados originais 33
Figura 4 – Otimização dos hiperparâmetros com LOGOCV 37
Figura 5 – Otimização dos modelos (conjunto de dados suíço) 38
Figura 6 – Otimização dos modelos (conjunto de dados agrupado) 38
Figura 7 – *Feature importance* (*Ada Boost*) 39
Figura 8 – *Feature importance* (conjunto de dados suíço) 40
Figura 9 – *Feature importance* (conjunto de dados agrupado) 41
Figura 10 – *Violin plot* da distribuição da classificação das *features* 42
Figura 11 – Teste U de Mann-Whitney da distribuição da classificação das features . . . 43
Figura 12 – *Violin plot* da distribuição da classificação das *features* (conjunto de dados suíço) 44
Figura 13 – Teste U de *Mann-Whitney* da distribuição da classificação das *features* (conjunto de dados suíço) 44
Figura 14 – *Violin plot* da distribuição da classificação das *features* (conjunto de dados agrupado) 45
Figura 15 – Teste U de *Mann-Whitney* da distribuição da classificação das *features* (conjunto de dados agrupado) 45

SUMÁRIO

1 INTRODUÇÃO 17
1.1 OBJETIVOS 17
1.1.1 Objetivo geral 17
1.1.2 Objetivos específicos 18
1.2 ORGANIZAÇÃO DO TRABALHO 18

2 FUNDAMENTAÇÃO TEÓRICA 19
2.1 LICITAÇÃO PÚBLICA 19
2.1.1 Modalidades de licitação pública 20
2.1.2 Conluio em licitações públicas 21
2.1.3 Técnicas de detecção de fraudes em licitações 21
2.1.3.1 Screening variables 21
2.2 *MACHINE LEARNING* 22
2.2.1 *Gradient Boosting* 22
2.2.2 *Adaptive Boosting* 23
2.2.3 *Extremely Randomized Trees* 23
2.2.4 *Multi Layer Perceptron* 23
2.2.5 *Logistic Regression* 24

3 TRABALHOS RELACIONADOS 25

4 METODOLOGIA 29
4.1 COLETA E DETALHAMENTO DOS CONJUNTOS DE DADOS 29
4.1.1 Conjunto de dados brasileiro 29
4.1.2 Conjunto de dados suíço 31
4.1.3 Conjunto de dados agrupado 32
4.2 REPRODUÇÃO DOS RESULTADOS 32
4.3 OTIMIZAÇÃO DE HIPERPARÂMETROS DOS MODELOS 34
4.4 ANÁLISE DOS RESULTADOS 34
4.5 APLICAÇÃO EM OUTROS CONJUNTOS DE DADOS 35

5 RESULTADOS E DISCUSSÃO 37
5.1 OTIMIZAÇÃO DE HIPERPARÂMETROS 37
5.1.1 Conjunto de dados brasileiro 37
5.1.2 Conjunto de dados suíço 38
5.1.3 Conjunto de dados agrupado 38
5.2 *FEATURE IMPORTANCE* 39
5.2.1 Conjunto de dados brasileiro 39

5.2.2 Conjunto de dados suíço 40
5.2.3 Conjunto de dados agrupado 40
5.3 ANÁLISE DA DISTRIBUIÇÃO DOS RESULTADOS 41
5.3.1 Conjunto de dados brasileiro 41
5.3.2 Conjunto de dados suíço 43
5.3.3 Conjunto de dados agrupado 45

6 CONCLUSÃO 47
6.1 TRABALHOS FUTUROS 47

REFERÊNCIAS 49

APÊNDICE A – INSTRUÇÕES PARA EXECUÇÃO DO FLUXO DE TRABALHO 51
A.1 UTILITÁRIOS 51

APÊNDICE B – ARTIGO SBC 53

1 INTRODUÇÃO

O processo de licitação consiste em uma ferramenta utilizada globalmente pelas administrações públicas para adquirir bens e serviços do setor privado. De acordo com Thorstensen (2021), estima-se que no Brasil, cerca de 12% do Produto Interno Bruto (PIB) tenha sido destinado a compras públicas entre os anos de 2002 a 2019, sendo a sua maioria por meio de processos licitatórios, o que demonstra ser um volume significativo de recursos públicos que passam por esse processo.

Embora existam leis destinadas a promover a livre concorrência entre os participantes, a prática de fraudes em licitações é uma ocorrência frequente e complexa de ser detectada pelos órgãos de controle. Essas fraudes podem assumir formas diretas, com empresas conspirando para fixar preços, ou indiretas, especialmente em mercados dominados por oligopólios, nos quais poucas empresas têm a capacidade de participar de licitações específicas. Isso resulta em uma série de problemas, como obras e serviços públicos de qualidade inferior para a população, redução da concorrência, dentre outros. (OECD, 2017).

Devido à abundância de dados gerados por esses processos nos últimos anos em forma digital, surgiram soluções que obtiveram êxito ao utilizar métodos avançados de análises de dados, como *machine learning*, para auxiliar na identificação de padrões fraudulentos. Dentre eles, está o trabalho de Velasco et al. (2021) onde foi implementado um sistema de suporte à decisão baseado em *machine learning* o qual auxiliou na investigação de contratos supostamente fraudulentos avaliados em cerca de R$ 3,6 bilhões.

Ainda que tais métodos tenham aperfeiçoado a identificação de fraudes, ainda existem desafios na escolha e ajuste de modelos para determinado caso. Anowar e Sadaoui (2019) elencam problemas como desbalanceamento de classes, similaridade de lances fraudulentos com comportamentos não fraudulentos e otimização de desempenho, que podem dificultar a análise do desempenho de modelos preditores.

Nesse sentido, este trabalho se propõe a analisar as atuais técnicas de detecção de fraudes em licitações utilizando dados disponibilizados publicamente a fim de auxiliar na decisão de quais modelos podem ser utilizados para determinado conjunto de dados, assim como apontar a necessidade de aperfeiçoá-lo. Para isso, serão utilizadas técnicas de otimização em modelos de *machine learning*, bem como a validação dos resultados das classificações realizadas por eles em diferentes conjuntos de dados.

1.1 OBJETIVOS

1.1.1 Objetivo geral

O objetivo geral deste trabalho consiste na análise da eficácia de modelos de *machine learning* para detectar fraudes em licitações públicas e a criação de um fluxo de trabalho que permita aprimorar o desempenho e facilitar a tomada de decisão para a escolha de qual modelo

utilizar em ambientes de produção.

1.1.2 Objetivos específicos

(i) Realizar uma revisão bibliográfica de técnicas de detecção de fraudes em licitações;

(ii) Reproduzir os experimentos encontrados na literatura;

(iii) Selecionar modelos com maior desempenho e realizar otimizações em cada um deles;

(iv) Avaliar o resultado do desempenho dos modelos otimizados em relação à literatura;

(v) Analisar a variabilidade dos resultados usando diferentes estratégias de divisão dos dados;

(vi) Analisar a importância das *features* do melhor modelo encontrado.

De forma geral, esta metodologia foi aplicada em três diferentes conjuntos de dados encontrados na literatura de base. O conjunto de dados de licitações de obras da Petrobras foi o primeiro a ser estudado. Além deste, foram realizados experimentos adicionais com o conjunto de dados suíço, que contém um número consideravelmente maior de registros em relação ao primeiro. E por fim, um experimento com todos os conjuntos do trabalho de referência agrupados. Foram reproduzidos os resultados, garantindo que sempre resultassem no mesmo valor, seguida pela otimização de modelos e comparação dos resultados, obtendo um ganho de desempenho de 8%, 11% e 5% nos respectivos conjuntos em relação ao trabalho base. Após essa etapa, os melhores modelos foram selecionados para realizar análises do comportamento das *features*. Nesse contexto, constatou-se a necessidade de enriquecimento em conjuntos menores e a importância da representatividade adequada para o problema em questão. Isso sugere que quanto mais específico e bem representado é o domínio, mais especializado é o modelo, gerando melhores resultados. No que diz respeito aos modelos, observou-se que o modelo *Logistic Regression* também pode apresentar resultados relevantes, dependendo da distribuição dos dados no conjunto, além de que modelos MLP mostraram um ganho de desempenho significativo mesmo com poucos dados, indicando que ainda há potencial para aperfeiçoamento.

1.2 ORGANIZAÇÃO DO TRABALHO

Os próximos capítulos deste trabalho estão organizados da seguinte forma: o Capítulo 2 consiste na fundamentação teórica, onde serão abordados conceitos fundamentais para a compreensão do trabalho, tais como licitações públicas, conluio, *machine learning* e diferentes modelos de *machine learning* que serão utilizados. O Capítulo 3 aborda os trabalhos relacionados. O Capítulo 4 explica a metodologia empregada para melhorar e avaliar os resultados. O Capítulo 5 expõe e discute sobre os resultados obtidos. Por fim, o Capítulo 6 apresenta as conclusões retiradas deste trabalho, como também apontar trabalhos futuros a serem realizados.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados conceitos fundamentais para a compreensão deste trabalho. A Seção 2.1 aborda a definição de licitação pública, suas modalidades, práticas de conluio e técnicas utilizadas para detecção. A Seção 2.2 apresenta a definição de *machine learning* e modelos utilizados, explicando resumidamente o seu funcionamento.

2.1 LICITAÇÃO PÚBLICA

De acordo com Pietro (2016), licitação pública consiste em um procedimento administrativo onde uma entidade pública abre a possibilidade para todos os interessados, ou licitantes, que estejam dispostos a respeitar os pré-requisitos estabelecidos em um instrumento convocatório, como um edital, para formularem propostas e, a partir delas, selecionar a mais conveniente para a celebração de um contrato.

No contexto da Lei n.º 14.133/21 (BRASIL, 2021), que estabelece as regras de licitação no Brasil, podemos entender a entidade pública citada por Pietro (2016) como a Administração Pública, sendo ela responsável pela administração direta ou indireta do bem público, bem como estatais e fundações instituídas ou mantidas por ela. Já os licitantes devem ser pessoas físicas ou jurídicas, ou um consórcio de pessoas jurídicas que participam ou manifestam o interesse em participar de uma licitação.

Vale ressaltar que esta lei substitui a Lei n.º 8.666/93 a fim de modernizar o ambiente de licitações públicas no Brasil. Até abril de 2023, era permitido ao gestor público escolher entre a anterior e a atual, porém sem mesclar ou alternar entre um regime e outro, dentre várias outras especificidades.

Além disso, a entidade pública é obrigada a utilizar este procedimento para adquirir bens e serviços, podendo haver a dispensa em casos específicos. No entanto, a regra é que cada processo licitatório passe pelas seguintes fases:

(i) Preparatória: ocorre dentro da entidade pública, os servidores fazem o levantamento do que deve ser adquirido e as regras do edital;

(ii) Divulgação do edital de licitação: o edital é publicado e disponível para ser acessado por qualquer indivíduo que esteja interessado;

(iii) Apresentação de propostas e lances: pode variar dependendo da modalidade. No caso da modalidade por concorrência, o licitante submete a sua proposta à entidade pública;

(iv) Julgamento: nesta etapa ocorre o julgamento do licitante vencedor, dependendo da modalidade e critérios adotados no edital;

(v) Habilitação: nesta etapa ocorre a verificação das documentações a fim de validar se o licitante possui a capacidade de realizar o que foi definido no edital em termos jurídicos, técnicos, fiscais, sociais, trabalhistas e financeiros;

(vi) Recursal: se algum dos licitantes que perderam na fase de julgamento não concordar com a decisão, deve-se elaborar um recurso embasando o motivo e enviar para a entidade pública, que irá analisar;

(vii) Homologação: se não houver algum impeditivo na fase recursal, é verificada a regularidade do processo licitatório, assim o processo é ratificado e aprovado.

Após finalizadas todas as etapas, a licitação é adjudicada, onde é atribuído oficialmente o vencedor da licitação e então realizada a assinatura do contrato. (BRASIL, 2021)

2.1.1 Modalidades de licitação pública

No Brasil, as licitações podem ser elaboradas na fase de preparação pelo gestor público conforme as seguintes modalidades: (BRASIL, 2021):

(i) Pregão;

(ii) Concorrência;

(iii) Concurso;

(iv) Leilão;

(v) Diálogo competitivo.

Como a modalidade de concorrência é a que está no escopo deste trabalho, será dado a esta um enfoque maior. Guimarães (2012 apud PENA, 2019) ressalta que esta modalidade é uma das mais amplas, uma vez que permite a participação de qualquer interessado que atenda aos requisitos fixados no edital e são utilizadas, em geral, para elaboração de contratos de grande porte, principalmente pelo fato de que a legislação anterior previa a obrigatoriedade dessa modalidade para alguns casos, como obras e serviços de engenharia de valor superior a R$ 1.500.000,00 e para compras e serviços que não sejam de engenharia, de valor superior a R$ 650.000,00 conforme evidenciado por Pietro (2016).

Também existem critérios definidos em Brasil (2021) que o gestor público poderá escolher usar na fase de julgamento da melhor proposta, sendo eles sintetizados por Junior (2022):

(i) Menor preço: O licitante vencedor será o que possuir a proposta com menor preço;

(ii) Melhor técnica ou conteúdo artístico: O julgamento é feito com base em negociações das condições ofertadas com o licitante melhor classificado, é geralmente utilizado para

serviços de caráter intelectual, como estudos técnicos preliminares ou gerenciamento, por exemplo;

(iii) Técnica e preço: O licitante vencedor é aquele que possui a maior média ponderada entre as propostas de preço e técnica;

(iv) Maior retorno econômico: É uma das modalidades mais recentes, ela consiste em um contrato de eficiência, onde o licitante vencedor deve apresentar na etapa de apresentação a maior economia estimada que pode gerar para a administração pública ao longo da execução do contrato;

(v) Maior desconto: O licitante vencedor será aquele que prover o maior desconto em relação ao preço fixado no edital.

2.1.2 Conluio em licitações públicas

De acordo com OECD (2017), conluio pode ser definido como qualquer prática de coordenação ou acordo entre empresas concorrentes em processos de licitação cujo objetivo seja maximizar os seus lucros, resultando em um prejuízo para os consumidores que, para este contexto, seria o pagador de impostos. Segundo Hendricks e Porter (1989), o conluio em licitações pode assumir muitas formas, dado a forma que os cartéis podem se adaptar caso a caso. Por exemplo, um cartel pode optar por revezar as vitórias entre si em várias licitações ou podem optar por dar lances adicionais após a vitória para dar a sensação de concorrência em um contexto de leilão.

Há também casos onde o próprio emissor da licitação ajuda determinado cartel a ganhar a licitação. Fazekas, Tóth e King (2013) comentam que esta técnica ocorre principalmente em organizações poderosas com grande influência política e que são capazes de obter vantagens, como ter acesso a informações relevantes da licitação, a fim de favorecê-las em relação às demais concorrentes.

2.1.3 Técnicas de detecção de fraudes em licitações

2.1.3.1 Screening variables

Screening variables são um conjunto de variáveis estatísticas baseadas na distribuição de lances e são utilizadas em análises estatísticas e econômicas para identificar possíveis comportamentos ilegais, como conluio, em determinado mercado, porém possuem um uso maior no contexto de licitações. Uma das grandes vantagens dessas variáveis é a facilidade de encontrá-las, uma vez que boa parte delas são obtidas por meio de dados públicos como preço, valor estimado da licitação, dados sobre o lance, participação de mercado, volume, dentre outros. (ABRANTES-METZ, 2013)

Essas variáveis se provaram como boas *features* em modelos de *machine learning* para detecção de fraudes, uma vez que conseguem evidenciar diferenças nas distribuições em relação às licitações puramente competitivas, como será apresentado posteriormente. Algumas das variáveis mais comuns são o coeficiente de variação (CV), a estatística de curtose (KURTO) e a dispersão (SPD) dos valores dos lances em uma licitação. (MITCHELL, 1997)

2.2 *MACHINE LEARNING*

Machine learning é um conjunto de técnicas que utiliza métodos computacionais e matemáticos para permitir que máquinas possam aprender a partir de dados e criar modelos especializados em fazer previsões ou realizar tomadas de decisões com base em novas observações. (ZHOU, 2021)

A fim de tornar o modelo mais preciso, existem diferentes abordagens dentro do escopo do *machine learning* que podem ser utilizadas, a depender do tipo de dado que será usado (texto, imagem, vídeo, áudio, dentre outros tipos de mídia), bem como a tarefa para a qual o modelo foi designado. Por exemplo, existem modelos especializados em realizar a classificação de determinada característica de um conjunto de dados, no caso de modelos de classificação, ou em calcular a probabilidade de pertencer a essa característica, no caso de modelos de regressão. (ZHOU, 2021)

Usualmente, em modelos preditores, o processo da construção do modelo envolve a divisão do conjunto de dados a ser usado em treinamento e teste, onde ambos possuem um conjunto de preditores (ou *features*) e as variáveis que se desejam prever (ou classes). Este primeiro conjunto de dados irá alimentar o modelo para poder encontrar padrões nos dados, ao passo que o conjunto de teste servirá para validar se o modelo foi capaz de generalizar para dados não vistos no treinamento. (ZHOU, 2021)

No entanto, existem alguns problemas com os quais essas abordagens precisam lidar, dois dos mais importantes são o viés e o *overfitting*: o primeiro pode ocorrer quando as classes são muito desbalanceadas entre si, fazendo com que o modelo não seja capaz de encontrar padrões para diferenciar corretamente, o que pode implicar uma queda no desempenho. Já o *overfitting* ocorre quando o modelo consegue aprender e prever com um bom desempenho, mas não é capaz de generalizar bem para novos dados. (ZHOU, 2021). A seguir, será apresentada a fundamentação teórica dos modelos de *machine learning* utilizados neste trabalho.

2.2.1 *Gradient Boosting*

De acordo com Natekin e Knoll (2013), as *Gradient Boosting Machines* (GBM) são uma família de técnicas de aprendizado de máquina com uma grande capacidade de capturar padrões onde não existe uma relação linear entre as variáveis.

A ideia geral por trás deste algoritmo está em realizar várias iterações de ajuste de um modelo simples, como uma árvore de decisão. Após o fim de uma iteração, são calculadas

as diferenças e as magnitudes dos erros em relação aos dados, a qual é chamada de gradiente negativo. Em seguida, é criado um novo modelo base ajustado para corrigir o erro do modelo da iteração anterior. Neste processo, cada modelo criado é adicionado ao conjunto de modelos (*ensemble*). Por fim, é calculada a previsão final somando todas as previsões dos modelos individuais a fim de que os erros se cancelem e o resultado seja uma previsão mais precisa usando o *ensemble*.

Além disso, os GBMs são altamente personalizáveis em questão de ajuste de hiperparâmetros, como a seleção da função de perda que será usada para calcular o erro residual de cada modelo individualmente, o que torna esta técnica vantajosa para diferentes aplicações.

2.2.2 *Adaptive Boosting*

Assim como o *Gradient Boosting*, o *Adaptive Boosting* (*Ada Boost*) também pertence à mesma família de modelos *ensemble*, que tem como princípio a combinação de "algoritmos fracos", para compor um algoritmo mais eficaz. No entanto, uma das diferenças entre eles é quanto à ênfase no ajuste dos modelos à medida que são treinados: o *Ada Boost* dá mais importância às instâncias que foram previstas incorretamente e atribui um peso com base na precisão. Após o treinamento dos modelos individualmente, é construído um classificador final que irá fazer a previsão usando os modelos individuais em função do peso dado para cada um deles. (MARGINEANTU; DIETTERICH, 1997)

2.2.3 *Extremely Randomized Trees*

Extremely Randomized Trees ou *Extra Trees* é uma variação de algoritmos baseados em árvores de decisão e *ensemble*. Basicamente, o algoritmo faz uma amostragem aleatória do conjunto de treinamento sem substituição para compor várias árvores de decisão, onde cada nó da árvore é dividido pelas *features* de forma aleatória, diferentemente de modelos similares que realizam esta divisão de maneira mais precisa com base em algum critério predefinido.

O objetivo geral deste algoritmo é ser computacionalmente mais eficiente para dados com muitas dimensões, uma vez que realiza a divisão dos nós de forma aleatória. Além disso, também pode ser indicado em casos de dados ruidosos ou suscetíveis a *overfitting* por fazer uma amostragem sem reposição. (GEURTS; ERNST; WEHENKEL, 2006)

2.2.4 *Multi Layer Perceptron*

Por sua vez, a rede Multi Layer Perceptron (MLP) compreende um tipo de rede neural artificial composta por uma camada de entrada de dados, uma ou mais camadas ocultas e uma camada de saída. Todos os nós, ou neurônios, estão totalmente conectados aos neurônios da camada seguinte e, para o caso dos neurônios ocultos, é introduzida uma função de ativação, como a função sigmoide.

O treinamento desse tipo de modelo ocorre através do algoritmo *backpropagation*, onde a cada iteração é realizado o ajuste dos pesos de cada conexão entre os neurônios utilizando uma função de perda, como entropia cruzada no caso de problemas de classificação. Em geral, este tipo de modelo é utilizado para problemas complexos, sistemas de carros autônomos, evolução de reações químicas, dentre outros. (RIEDMILLER; LERNEN, 2014)

2.2.5 *Logistic Regression*

Por fim, a regressão logística, ou *logistic regression*, é uma técnica estatística amplamente utilizada em classificações binárias. O algoritmo é baseado em uma função logística sigmoide ajustada em função da combinação linear das características, ponderadas pelos coeficientes aplicados em cada uma delas, ou seja, coeficientes com maior peso têm uma maior importância para a função logística.

Durante o treinamento, esses coeficientes são ajustados a fim de reduzir a função de custo, utilizando técnicas para otimizá-los, como gradiente descendente. Após isso, é definido um limiar de decisão e, caso uma observação esteja dentro desse limiar, é atribuída à classe positiva, caso contrário, é atribuída à classe negativa.

3 TRABALHOS RELACIONADOS

Tendo em vista os conceitos que envolvem um processo licitatório e a complexidade na detecção de fraudes citadas anteriormente, este capítulo retrata as diferentes técnicas de *machine learning* utilizadas pelo estado da arte para identificação de conluio entre licitantes.

No artigo *"Detection of bid rigging in procurement auctions"*, Porter e Zona (1993) fazem uma análise sobre contratos de construção de rodovias nos condados de Nassau e Suffolk, em Nova Iorque, entre abril de 1979 e março de 1985, a fim de propor uma metodologia baseada em testes estatísticos e econométricos, que são uma extensão do uso de ferramentas estatísticas voltadas para modelos econômicos, para detectar a presença de conluio entre as empresas participantes das licitações.

Neste estudo, foi concluído que o comportamento de lances e valores de empresas competitivas possuem diferenças estatísticas em relação aos cartéis. Por exemplo, ao aplicarem o teste de Chow, que consiste em dividir o conjunto de dados entre colusivos e não colusivos para validar se há uma diferença abrupta em determinadas variáveis aleatórias, resultou na hipótese de não existir competidores colusivos, descartando esta hipótese.

No entanto, os próprios autores reconheceram haver limitações em seus estudos, uma vez que existe a possibilidade de terem classificado inadvertidamente algumas empresas do cartel como competitivas, bem como a falta de variáveis explicativas. Ademais, se alguma agência anunciasse publicamente o uso desta metodologia, haveria um risco de os cartéis se valerem da mesma para contornarem esta detecção, aumentando proporcionalmente o valor dos seus lances, por exemplo.

Com o avanço no poder de processamento de dados e *data mining*, Velasco et al. (2021) propuseram no artigo *"A decision support system for fraud detection in public procurement"* um sistema para auxiliar órgãos de monitoramento com indícios de fraudes em licitações. Para isso, foi proposto um sistema de suporte à decisão que utiliza um conjunto de técnicas de extração de dados de diferentes fontes.

Diferentemente de Porter e Zona (1993) que utilizaram métodos já consolidados em econometria, os próprios autores elaboraram um conjunto de comportamentos que poderiam indicar uma prática fraudulenta na licitação, denominado *risk patterns*. Quanto mais desses padrões de risco existirem, maior a probabilidade da empresa estar envolvida em um esquema de fraude. Dentre os mais de 20 *risk patterns* e 200 variáveis usadas para computá-las, uma delas é a quantidade de vezes que uma empresa estava envolvida em um processo de licitação enquanto já havia acusações de ter participado de algum esquema de fraude.

Esse estudo ressalta o uso da metodologia em casos operacionais, como investigação de empresas fictícias envolvidas em processos de licitação, bem como dados quantitativos do impacto para diferentes tipos de fraudes. Uma delas foi a análise de 857 empresas que venceram contratos de licitação, mas que possuíam empresas em comum disputando o mesmo edital, somando uma quantia maior do que R$ 3,6 bilhões em contratos.

Os autores também ressaltam que esta metodologia não exclui uma análise manual,

a qual eles julgam ser muito importante para analisar todas as informações produzidas pelo modelo, assim como reconhecem que este possui limitações, como a necessidade de uma alta quantidade e qualidade de dados, uma vez que sem esses requisitos, a tomada de decisão pode ser diretamente afetada.

Por sua vez, no artigo *"Prediction of public procurement corruption indices using machine learning methods"*, Rabuzin e Modrusan (2019) também utilizam técnicas de *data mining* para extração de documentações sobre licitação em órgãos públicos da Croácia. Para isso, os autores utilizaram um modelo baseado em processamento de linguagem natural (MLP) para extrair variáveis como condições técnicas, prazo, valores estimados, dentre outros.

A partir disso, essas variáveis foram utilizadas para treinar diferentes modelos de *machine learning*, sendo eles *support vector machine*, *naive bayes* e regressão logística. Neste artigo, podemos notar que os autores também utilizaram técnicas de *data mining* para extração de informações sobre licitação, com a diferença que, em vez de utilizar variáveis predefinidas, os dados foram inseridos em modelos de *machine learning*, de modo a fazer com que o processo de descoberta dos padrões fosse realizado pelos próprios modelos.

Para avaliar o desempenho dos modelos, foram avaliadas métricas de acurácia, precisão, *recall* e ROC, tanto para a validação utilizando todo o conjunto de dados quanto para dados agrupados por categoria de licitação. A precisão das previsões variou entre 60% e 85%, já a métrica de *recall* teve um resultado acima de 5%, sendo considerado um bom resultado pelos autores.

Os autores também encontraram limitações em seus estudos, dentre elas estão o fato de usarem um conjunto de dados muito pequeno, uma vez que os órgãos reguladores croatas não divulgam muitos dados se determinado lance era de uma empresa colusiva, reduzindo a quantidade de fontes acessíveis para usar nos modelos. Além disso, os autores ressaltam que a falta de preditores pode ter influenciado no desempenho do modelo, fazendo com que surgissem variáveis adicionais, bem como o uso de modelos mais sofisticados, como redes neurais e aprendizado profundo (*deep learning*) para melhorar os resultados.

Devido ao surgimento de soluções utilizando *machine learning* para detectar fraudes em licitações, Rodríguez et al. (2022) realizaram um estudo transversal para testar o desempenho de diferentes combinações de modelos, conjuntos de dados e configuração de *features*. A metodologia consistiu na coleta de dados de contratos de licitação envolvidos em fraudes de diferentes países, dentre eles, os dados de licitações investigados pela operação Lava Jato em 2014. Após a coleta, foram selecionados e treinados 11 modelos em diferentes configurações de dados, incluindo o uso de *screening variables*.

Diferentemente dos demais autores citados anteriormente, que buscaram realizar um estudo para um conjunto de dados em específico, estes propuseram um estudo que abrangesse diferentes condições de dados, algoritmos e configurações de *features*, sendo o primeiro estudo a ter essa abordagem. Após o treinamento e validação dos resultados, foi realizada a análise usando métricas como acurácia balanceada, onde obteve um desempenho normalmente acima dos 80% e falsos positivos e negativos abaixo de 10%. Apesar disso, os autores também en-

contram limitações no estudo, onde ressaltam uma análise mais aprofundada de otimização de hiperparâmetros, uso de mais variáveis e dados para auxiliar no bom desempenho dos modelos.

Por fim, no artigo *"A machine learning approach for flagging incomplete bid rigging cartels"*, Wallimann, Imhof e Huber (2022) realizaram o uso de técnicas de regressão utilizando *machine learning* em conjuntos de dados de fraudes em licitações nas regiões de See-Gaster e Graubünden, na Suíça, a fim de estimar a probabilidade de cartéis estarem participando de licitações através dos lances onde nem todos os concorrentes da licitação estavam envolvidos no esquema.

De forma similar a Rodríguez et al. (2022), os autores também utilizaram *screenings variables* para auxiliar os modelos na classificação, agrupando as amostras em diferentes subgrupos por licitação para calcular essas variáveis. Após isso, foi realizado o treinamento com diferentes configurações de cálculo das *screenings* em um modelo *Random Forest*. Os autores também utilizaram técnicas de *feature importance* para descartar possíveis variáveis que não possuíam importância significativa para o desempenho do modelo.

Após a validação, foram comparadas as taxas de classificação correta em relação a um modelo de referência para diferentes amostras e configurações das *features*, a qual variou entre 61,2% e 84,1%. Os autores também ressaltaram que esta diferença se deu por conta de supostos casos onde o cartel eventualmente não praticava o conluio total em determinada licitação. Além disso, o uso de configurações de amostras onde havia o valor do lance e a quantidade de licitantes como preditores aumentaram as taxas de classificação na ordem de 5 a 10 pontos percentuais em relação aos demais.

A partir dos trabalhos relacionados, o presente trabalho tem como base a pesquisa realizada por Rodríguez et al. (2022), com o intuito de aperfeiçoar o fluxo proposto para garantir que a reprodução dos resultados seja perfeitamente replicável, somado a um processo de otimização, seleção e análise dos modelos, observando o desempenho dos algoritmos com base no comportamento das *features* utilizadas no processo de treinamento. Dessa forma, é possível analisar não somente o resultado com base em métricas, como também uma análise visual do comportamento das *features*, o que pode contribuir para a tomada de decisão do melhor modelo ou se existe a necessidade de alguma modificação no conjunto de dados utilizado.

4 METODOLOGIA

A metodologia deste trabalho consiste em realizar análises sobre o conjunto de dados disponibilizado no estudo de Rodríguez et al. (2022), aplicar otimizações sobre os modelos com melhor desempenho evidenciado pelos autores, além de testar modelos adicionais e, por fim, analisar os resultados obtidos. A motivação para a utilização deste conjunto em específico se dá pela riqueza de informações sobre diferentes casos de fraudes em licitações em diversos países. Além disso, trata-se de casos de fraudes recentes e com informações amplamente divulgadas, facilitando a compreensão dos dados.

A Figura 1 ilustra o fluxo aplicado neste e em outros dois conjuntos de dados desse mesmo artigo de referência. No passo 1, os dados foram coletados do material suplementar fornecido pelo artigo de referência. No passo 2, foram reproduzidos os algoritmos fornecidos pelo artigo. No passo 3, foram aplicadas técnicas para garantir a replicabilidade dos dados. No passo 4, foram designadas as estratégias de divisão dos dados com base nos conjuntos utilizados. No passo 5, foram otimizados os hiperparâmetros dos modelos para cada conjunto de dados, visando um aumento no desempenho em relação aos resultados originais. No passo 6, foram selecionados os melhores modelos de cada conjunto com base nos resultados da otimização. No passo 7, foram analisadas a importância das *features*. No passo 8, foram analisadas a distribuição das classificações corretas das incorretas. No passo 9, foram utilizados testes estatísticos das distribuições classificadas corretamente das incorretas. Por fim, no passo 10, serão apresentadas as conclusões dos experimentos realizados, assim como sugestões de trabalhos futuros.

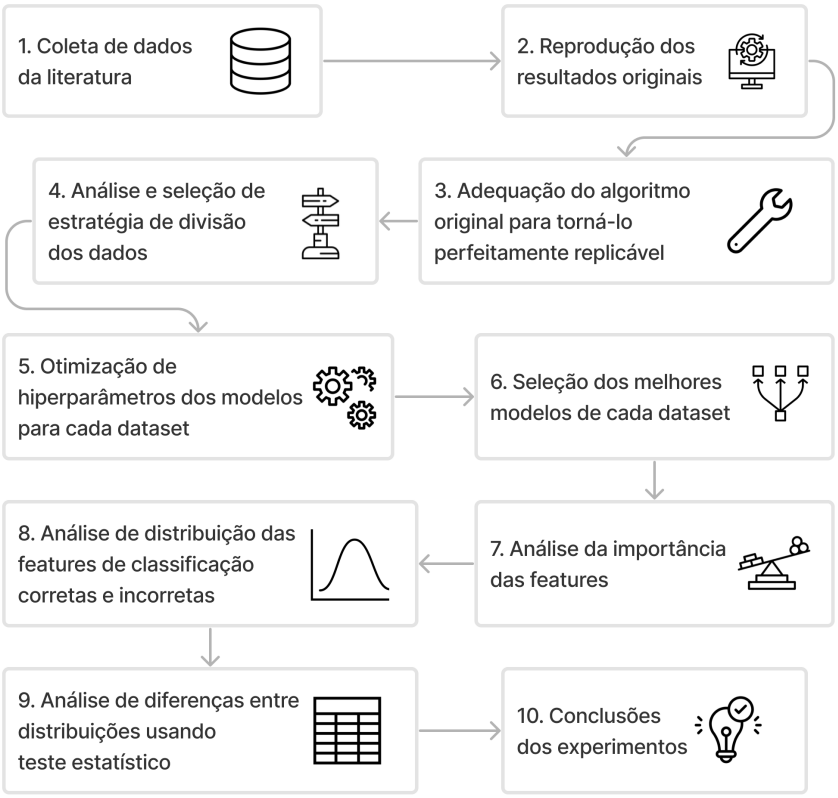
4.1 COLETA E DETALHAMENTO DOS CONJUNTOS DE DADOS

Nesta primeira etapa do fluxo de trabalho, os conjuntos de dados utilizados e o código original foram obtidos através do material suplementar fornecido no artigo de Rodríguez et al. (2022). Sendo essa a primeira etapa no fluxo de trabalho, apresentado na Figura 1. Dentre eles, o conjunto de conluio em licitações de obras da empresa Petróleo Brasileiro (Petrobras) foi o objeto do estudo mais aprofundado. Além deste, foram realizados experimentos adicionais com o conjunto de dados suíço, que contém um número consideravelmente maior de registros. E por fim, um experimento com todos os conjuntos do trabalho de referência agrupados. A seguir, tais conjuntos serão apresentados e detalhados.

4.1.1 Conjunto de dados brasileiro

A operação Laja Jato foi uma operação conduzida pela Polícia Federal em 2014 que investigou um esquema de lavagem de dinheiro, onde um dos investigados estaria ligado a um ex-diretor da Petrobras.

Figura 1 – Fluxograma da metodologia realizada neste trabalho



Por sua vez, a Petrobras é uma empresa de capital aberto, com controle majoritariamente estatal do Brasil, cuja operação consiste em realizar a extração, refino e transporte de petróleo e gás no país. Após sucessivos desdobramentos, esse ex-diretor estaria supostamente ligado a um cartel onde receberia subornos em troca de contratos para obras da companhia. Em 2015, a empresa divulgou em seu relatório anual aos acionistas que perdeu quase 6 bilhões de reais devido a esquemas de corrupção, equivalente a 1,9 bilhão de dólares naquele período. (MORO, 2018)

Neste conjunto de dados, existem 683 lances com 272 licitantes, distribuídos em 101 licitações de contratos de obras e serviços. Para cada lance, o conjunto possui as seguintes informações:

- *Tender*: Identificador da licitação;
- *Bid*: Identificador do lance na licitação;
- *Competitors*: Identificador do licitante;
- *Bid Value*: Valor do lance;
- *Pre-Tender Estimate* (PTE): Estimativa do valor da licitação fixado no edital;
- *Date*: Data do lance;
- *Site*: Local onde a obra ou serviço foi executado;

- *Brazilian State*: Identificador do estado onde ocorreu a licitação;
- *Difference Bid/PTE*: Diferença do lance para a estimativa do valor da licitação;
- *Number Bids*: Número de lances que ocorreram em determinada licitação;
- *Winner*: Identificador da empresa vencedora da licitação;
- CV (Coefficient of Variation): Variabilidade dos lances em relação à média;
- *Spread* (SPD): Dispersão do valor dos lances na licitação;
- *Relative Difference* (DIFFP): Diferença relativa entre os dois lances mais baixos em uma licitação;
- *Relative Distance* (RD): Desvio padrão entre os dois lances mais baixos em uma licitação;
- *Excess Kurtosis*(KURT): Grau de achatamento ou picos nas distribuições dos lances, usada apenas em licitações com mais de quatro lances;
- *Skewness* (SKEW): Grau de assimetria na distribuição dos lances;
- *Kolmogorov-Smirnov test* (KSTEST): Grau de similaridade dos valores dos lances em uma distribuição uniforme;
- *Collusive competitor*: Booleano que indica se a licitação havia conluio (considera-se conluio quando mais de 11% dos lances eram colusivos);
- *Collusive competitor original*: Booleano que indica se o lance foi originado de um coluio;

Os campos CV, SPD, DIFFP, RD, KURT, SKEW, KSTEST se referem às *screening variables* utilizadas pelos autores para aumentar o desempenho dos modelos. Outro ponto importante é que para todos os conjuntos de dados estudados, a variável de saída foi a *Collusive competitor*.

4.1.2 Conjunto de dados suíço

O conjunto de dados suíço representa os dados obtidos por meio de um esquema de conluio em licitações de infraestrutura rodoviária nos cantões suíços de St. Gallen e Graubünden (SG&GR) entre os anos de 2004 a 2010. No primeiro cantão, oito empresas se reuniam uma ou duas vezes por mês para negociar lances de licitações. Já no segundo, uma associação comercial era responsável por fraudar licitações de obras de estradas e pavimentação asfáltica. Após a descoberta desses esquemas, a *Swiss Competition Commission* (COMCO) realizou buscas e apreensões domiciliares nas residências dos responsáveis por esses cartéis. (WALLIMANN; IMHOF; HUBER, 2022)

Esse conjunto de dados possui 4.344 licitações e 21.231 lances. A motivação para a escolha deste se deu pela quantidade e qualidade dos dados, uma vez que, diferentemente do conjunto de dados agrupado, os dados são estruturados para apenas um escopo de fraude, além de contar com as seguintes *features*: *Tender*, *Bid Value*, *Winner*, *Number Bids*, *Date*, *Contract type* e *Collusive competitor*.

4.1.3 Conjunto de dados agrupado

Outro conjunto de dados objeto desse estudo foi o que reúne todos os outros conjuntos estudados pelo artigo de referência em apenas um. Esse conjunto possui 9.781 licitações e 64.348 lances compilados em seis conjuntos diferentes. A escolha deste conjunto se deu pela quantidade de licitações e lances, algo que o conjunto de dados brasileiro carece. Contudo, a quantidade de *features* utilizadas foi limitada àquelas utilizadas pelo autor, a fim de ser possível comparar os resultados, sendo elas *Tender*, *Bid Value*, *Winner*, *Number Bids* e *dataset*. Esta última *feature* corresponde ao identificador do conjunto de dados de onde foi extraído o dado.

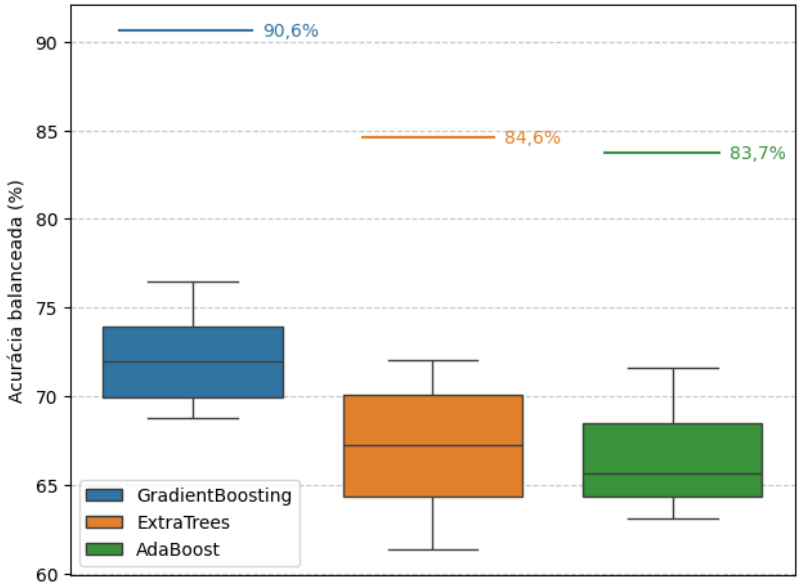
Embora a *feature* "*Dataset*" seja utilizada pelo artigo de referência, mas não seja tangível em um conjunto de dados utilizado em produção, a ideia é verificar se os modelos conseguem generalizar mesmo com uma variedade de dados tão grande, a medida que são subconjuntos de dados de diferentes contextos de fraudes.

4.2 REPRODUÇÃO DOS RESULTADOS

Nesta segunda etapa do fluxo de trabalho, foi constatado que, no código original, os resultados foram obtidos por meio de 50 iterações de treinamento e teste dos modelos. O desempenho consistiu na média dessas iterações, técnica esta chamada de *random subsampling* (TAN; STEINBACH; KUMAR, 2016). Todavia, não foram reportados os desvios padrão, e nem fixadas, no código disponibilizado, as sementes para geração de números aleatórios, de forma que a reprodução exata dos resultados pudesse ser realizada.

A Figura 2 apresenta a acurácia balanceada obtida neste estudo após 10 execuções do algoritmo original para os três melhores modelos. As linhas com as respectivas cores nas legendas correspondem aos resultados reportados pelos autores. Como pode ser observado, a acurácia balanceada do artigo é na ordem de 10 a 20% acima da encontrada após a reprodução dos resultados neste estudo. Em teoria, a tendência geral dos resultados deveria se manter mesmo com diferentes *seeds*. No entanto, devido à ausência de informações sobre o desvio padrão no artigo de referência, não haveria como confirmar se essa variabilidade seria aceitável no desvio padrão obtido pelos autores.

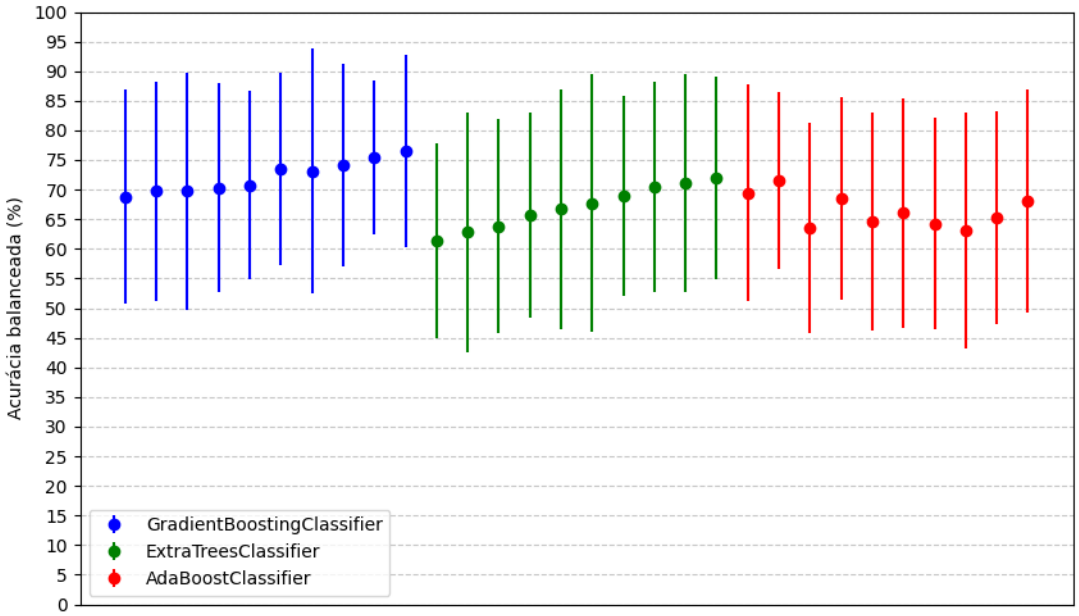
Figura 2 – Boxplot da reprodução dos resultados originais



Fonte: Autor

Para ilustrar melhor os resultados obtidos ao executar o algoritmo original, a Figura 3 apresenta a acurácia balanceada e o desvio padrão dos modelos em cada uma das 10 execuções utilizando *random subsampling*. Podemos notar uma diferença próxima na ordem de 10% entre as iterações, bem como um desvio padrão de 15% a 20%.

Figura 3 – Barra de erro da reprodução dos resultados originais



Fonte: Autor

Vale ressaltar que para o treinamento do modelo, foram utilizadas as *features Tender*, *Bid Value*, *Winner*, *Number Bids*, *PTE*, *Difference Bid/PTE*, *Site*, *Brazilian State* e *Date*, assim como as *screening variables*.

Dada a grande variabilidade dos resultados e a quantidade limitada de dados neste conjunto de dados, foi utilizada a técnica *Leave One Group Out Cross Validation* (LOGOCV). Essa técnica é uma forma especial de validação cruzada que se assemelha ao método *k-fold*, com a diferença de que, no LOGOCV, o valor de *k* é igual ao número de grupos. Em cada iteração, dos *k* grupos, todos os grupos, exceto um, são usados para treinamento, e o grupo excluído é usado para teste. Neste caso, cada iteração do LOGOCV treina o modelo *k* - 1 *Tenders*. A decisão por agrupar por *Tender* seguiu a mesma implementação do artigo de referência, evitando que lances de determinada licitação estivessem em ambas as divisões, contaminando o conjunto de treinamento.

De acordo com Wong (2015), o uso do LOGOCV é comumente aplicado em conjuntos de dados pequenos, uma vez que exclui a possibilidade de haver variabilidade na divisão aleatória dos dados em treinamento e teste, pois todas as possíveis divisões são consideradas, obtendo um resultado sempre constante.

4.3 OTIMIZAÇÃO DE HIPERPARÂMETROS DOS MODELOS

Na quinta etapa do fluxo de trabalho, foram realizadas otimizações nos 3 melhores modelos para as diferentes configurações e em outros 2, a fim de comparar com os resultados dos demais, sendo eles o *Multi Layer Perceptron* (MLP) e *Logistic Regression*.

O MLP também foi estudado pelo artigo de referência, porém não atingiu os melhores resultados. Todavia, como se trata de um modelo amplamente utilizado em estudos envolvendo classificação, foi sujeito ao processo de otimização com o propósito de se obter melhores resultados. Já a escolha do *Logistic Regression* se deu por ser um modelo fundamental no contexto de técnicas de classificação e servirá de critério para determinar um limite inferior dos resultados obtidos pelos demais modelos.

Juntamente com a técnica LOGOCV, foi utilizado também o *Grid Search*, que constitui em uma automatização do processo de busca de hiperparâmetros onde é definido um espaço de valores de diferentes hiperparâmetros e então combinados para realizar o treinamento do modelo. Após encontrado, o modelo é retreinado a partir de uma técnica de validação cruzada predefinida pelo programador.

4.4 ANÁLISE DOS RESULTADOS

Na sétima etapa do fluxo de trabalho, após a seleção dos melhores modelos, foram utilizados métodos para analisar os resultados, de modo que fosse possível extrair conhecimento adicional do processo de treinamento e teste dos modelos, evidenciando as características mais importantes utilizadas por eles para a detecção de conluio. Para isso, foi extraída a *feature importance* dos melhores modelos de cada conjunto de dados. Essa métrica compreende a importância média que cada *feature* teve na utilização para a realização da detecção por parte dos modelos. A interpretação destes valores por especialistas do domínio do problema, isto é,

promotores, auditores, fiscais e etc., tem especial valor para aprimorar a prevenção e detecção de fraudes em licitações públicas.

Além disso, de forma a melhor entender os erros de classificação dos melhores modelos, as licitações foram divididas entre as classificadas correta e incorretamente, e posteriormente foram analisadas as distribuições dos valores de cada *feature* dessas classificações. Para isso, tais distribuições foram ilustradas em gráficos *violin plots* para cada *feature* numérica utilizada nos modelos. Esse tipo de gráfico combina elementos do gráfico de *boxplot* com um gráfico de densidade de probabilidade.

A partir da distribuição das classificações corretas e incorretas de cada *feature*, foram realizados testes estatísticos para identificar a influência de cada *feature* na classificação final. Esse tipo de análise tem por objetivo verificar, com um intervalo de confiança de 95%, se existe diferença entre as duas distribuições. Por exemplo, se uma *feature* com maior importância possuir uma diferença nas suas distribuições, então pode-se afirmar com 95% de confiança que essa *feature* contribuiu para o bom ou mau desempenho do modelo. Este estudo é particularmente relevante considerando o possível impacto do conjunto de dados ser limitado a poucos dados, de forma que os valores das *features* podem ter distribuições diferentes para os mesmos valores da variável-alvo. Para isso, todas as amostras foram inicialmente submetidas ao teste de *Shapiro-Wilk*, cujo objetivo é comparar se essas distribuições seguem uma distribuição normal. Caso afirmativo, são analisadas a partir do teste t de *Student*. Do contrário, são comparadas a partir do teste U de *Mann-Whitney*.

4.5 APLICAÇÃO EM OUTROS CONJUNTOS DE DADOS

Após consolidada a metodologia de análise para o *conjunto de dados* brasileiro, a mesma foi aplicada nos outros dois conjuntos disponibilizados no material suplementar do artigo de referência, demonstrando que a mesma sequência de análise pode ser realizada para qualquer conjunto de dados, além de verificar o desempenho dos modelos a partir de premissas que o conjunto de dados brasileiro não possui, como uma quantidade maior de lances e licitações.

Dado que os demais conjuntos de dados possuem mais licitações do que o conjunto de dados brasileiro, a opção de aplicar o LOGOCV nesses conjuntos seria computacionalmente custosa. Dessa forma, foi realizado o treinamento dos modelos com 50 *folds*, agrupados pela licitação, para ser possível comparar a acurácia balanceada. Em geral, quanto mais *folds*, mais próximo da realidade é o resultado, sendo o LOGOCV o caso mais “extremo” dessa técnica.

5 RESULTADOS E DISCUSSÃO

Nesta seção apresentamos os resultados experimentais obtidos ao aplicar o fluxo de trabalho proposto na metodologia para analisar os modelos treinados utilizando o conjunto de dados brasileiro com LOGOCV e os outros dois conjuntos com *GroupKFold* com 50 *folds*.

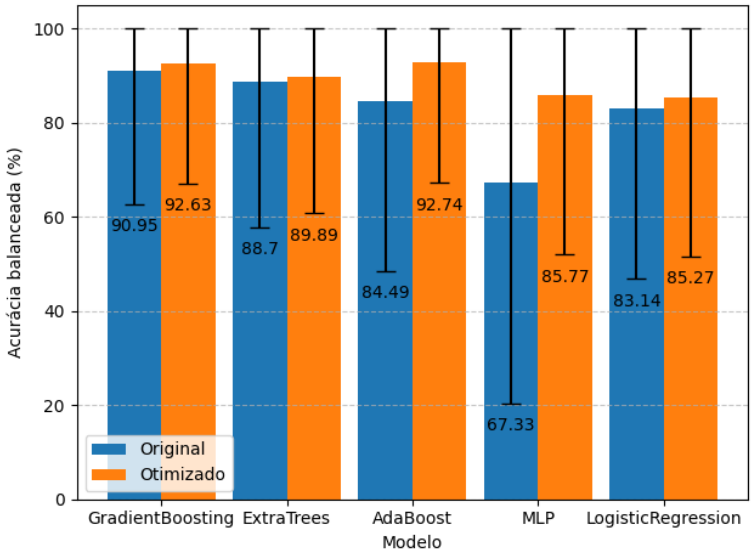
5.1 OTIMIZAÇÃO DE HIPERPARÂMETROS

5.1.1 Conjunto de dados brasileiro

A Figura 4 apresenta os resultados da aplicação do LOGOCV usando os mesmos hiperparâmetros do artigo de referência e comparando com os resultados da otimização deste estudo. Observa-se um ganho geral de desempenho em todos os modelos, sendo o *Ada Boost* o melhor, com 92,74%, ficando próximo ao resultado do *Gradient Boosting*, com 92,63%. Além disso, o MLP foi o que obteve o maior ganho de desempenho, aumentando sua acurácia balanceada em 18,44%. Todavia, nota-se que o desvio padrão ainda continua alto, o que demonstra que mesmo com a otimização dos resultados, os modelos ainda não possuem uma constância na predição das licitações.

Complementarmente, o *Logistic Regression* teve um ganho de desempenho de 2% apenas, sugerindo que a distribuição dos lances possui características não lineares, fazendo com que modelos mais robustos como *Ada Boost* consigam extrair essas características com maior acurácia. Também podemos verificar que os modelos com o mesmo conjunto de hiperparâmetros utilizados pelo artigo de referência obtiveram um desempenho melhor com o LOGOCV do que com *random subsampling* apresentado na Figura 2, ficando próximo do resultado obtido pelos autores.

Figura 4 – Otimização dos hiperparâmetros com LOGOCV



Fonte: Autor

5.1.2 Conjunto de dados suíço

De acordo com a Figura 5, podemos observar que o conjunto de dados suíço teve o *Logistic Regression* como o melhor modelo, com 83,54%. Apesar de ser considerado o modelo mais simples, também foi o que obteve o menor desvio padrão e o maior ganho de desempenho dentre os demais modelos. Isso pode ser explicado pelo fato de que as *features* com maior importância tendem a ser mais facilmente separadas por meio de uma função logística, como veremos adiante. Outro ponto observado foi que, como se trata de um modelo mais rápido de ser treinado, foi possível explorar um conjunto maior de hiperparâmetros em um tempo computacionalmente razoável, o que pode ter contribuído para obtenção de um conjunto que se adequasse mais às características do conjunto de dados em questão.

5.1.3 Conjunto de dados agrupado

Já para o conjunto de dados agrupado, observa-se a partir da Figura 6, um pequeno ganho de desempenho, sendo o *Gradient Boosting* o melhor, com 84,74%. Da mesma forma que o conjunto de dados brasileiro, observou-se uma estabilização nos resultados do *Logistic Regression*, indicando que podem existir características nos dados que não são lineares, fazendo com que o *Gradient Boosting* consiga explorá-las de forma mais eficiente.

Figura 5 – Otimização dos modelos (conjunto de dados suíço)

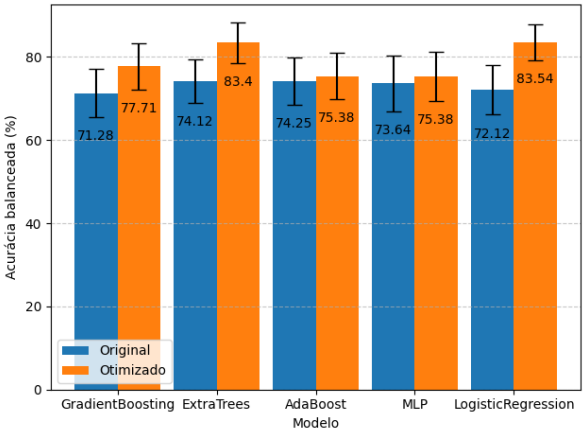
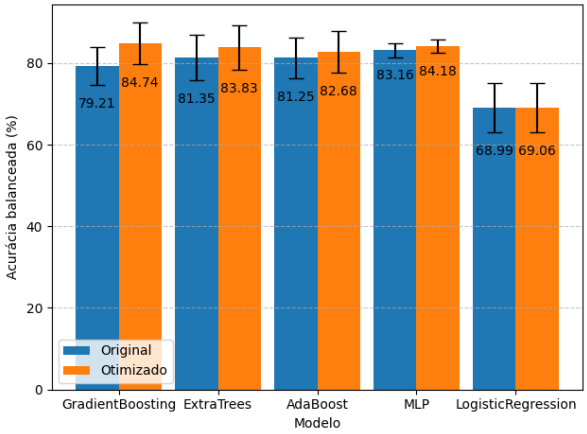


Figura 6 – Otimização dos modelos (conjunto de dados agrupado)

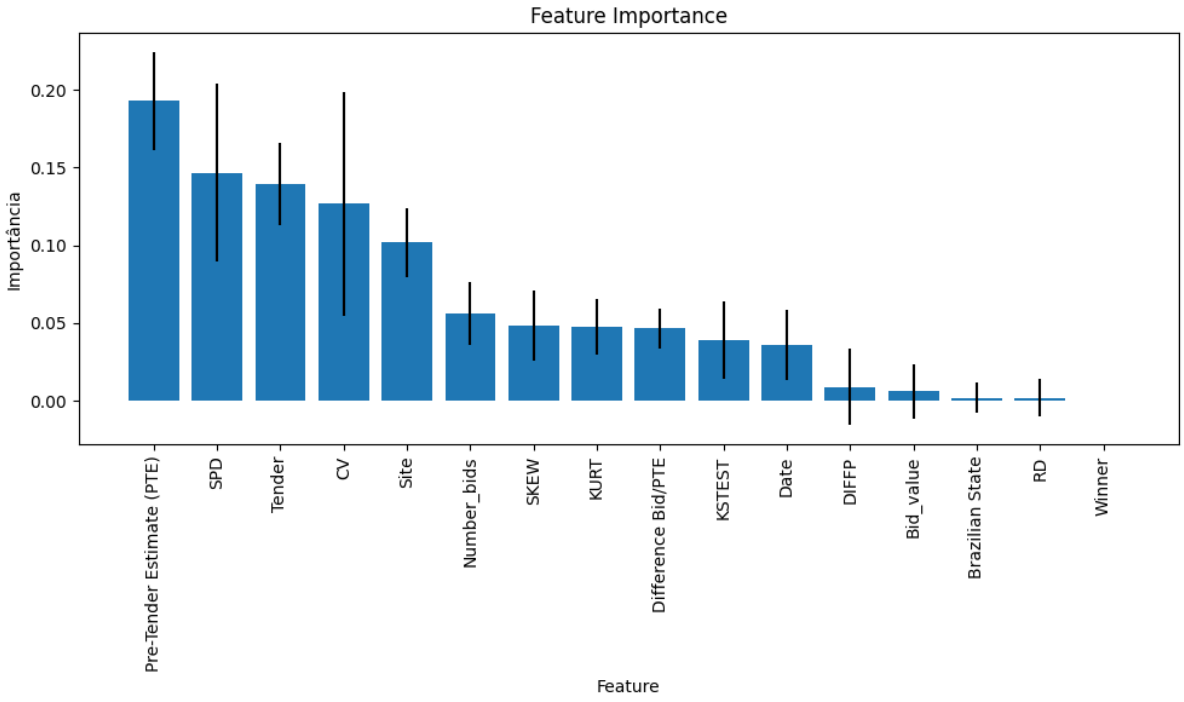


5.2 *FEATURE IMPORTANCE*

5.2.1 Conjunto de dados brasileiro

Após selecionado o melhor modelo, foram realizadas algumas análises a fim de compreender os resultados obtidos, dentre elas está a importância média das *features*. A Figura 7 representa a *feature importance* do *Ada Boost*.

Figura 7 – *Feature importance* (*Ada Boost*)



Fonte: Autor

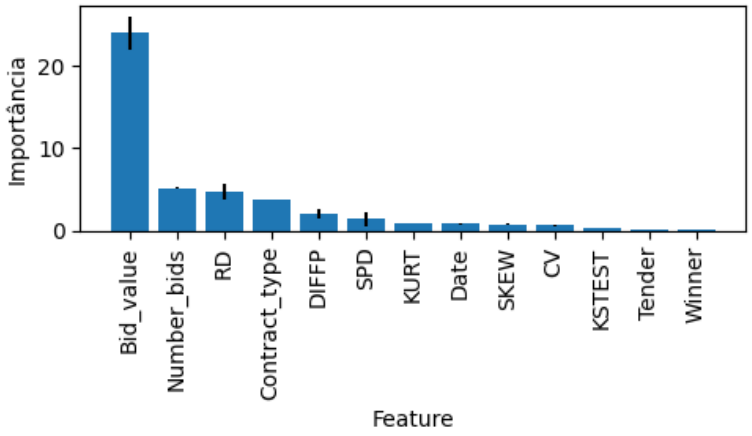
Examinando a figura, podemos verificar que o modelo utilizou *features* de diferentes categorias de forma distribuída, sem concentrar as importâncias em um conjunto reduzido. As mais importantes são PTE, SPD, *Tender*, CV e *Site*. No caso do PTE, podemos deduzir a partir da importância dessa *feature* que o potencial de fraudes é maior em licitações onde os lances são mais altos. Em outras palavras, as empresas colusivas tendem a focar em licitações de maior valor. Já o SPD e CV com uma maior importância indicam que a variabilidade nos lances pode estar associada a comportamentos colusivos. Tal situação pode ocorrer em casos de conluio, onde é acordado antecipadamente que determinada empresa irá oferecer um preço muito mais baixo para ganhar a licitação, criando uma falsa sensação de concorrência. No caso das *features* categóricas *Tender* e *Site*, constata-se que essa importância pode ser explicada pela frequência em que determinado identificador pertencente à classe colusiva aparece no conjunto de dados. Por exemplo, a *feature Site* com uma importância significativa indica que pode haver localizações com maior tendência de ocorrer conluio entre empresas.

5.2.2 Conjunto de dados suíço

Com os melhores modelos de cada conjunto de dados definidos, foi realizada a análise da importância das features de cada um, conforme representado nas Figuras 8 e 9.

Na primeira figura, nota-se que a *feature Bid Value* foi a que obteve um maior peso para a criação da curva de decisão do modelo *Logistic Regression*, com 23% de importância, seguida da *feature Number Bids*, com 5%. Ao observar esse resultado, podemos afirmar que o modelo conseguiu traçar uma curva e delimitar lances colusivos dos não colusivos, considerando essas *features* com um peso significativo. Todavia, embora exista essa relação, veremos nas distribuições dos resultados que essa mesma característica afetou também a taxa de erro do modelo.

Figura 8 – *Feature importance* (conjunto de dados suíço)

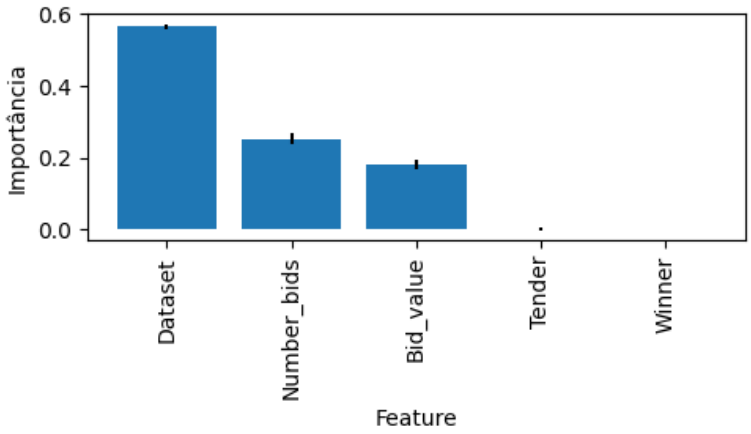


Fonte: Autor

5.2.3 Conjunto de dados agrupado

No conjunto de dados agrupado, podemos observar que o modelo *Gradient Boosting* usou a *feature* categórica *Dataset*, com 57% de importância, seguida também pelas *features Number Bids* e *Bid Value*, com 23% e 16% de importância, respectivamente. Podemos notar a partir disso que ambas as *features* possuem uma importância significativa em ambos os conjuntos, sugerindo a existência de valores de lances e quantidade de licitantes que podem contribuir para indicar a presença de conluio em determinada licitação.

Figura 9 – *Feature importance* (conjunto de dados agrupado)



Fonte: Autor

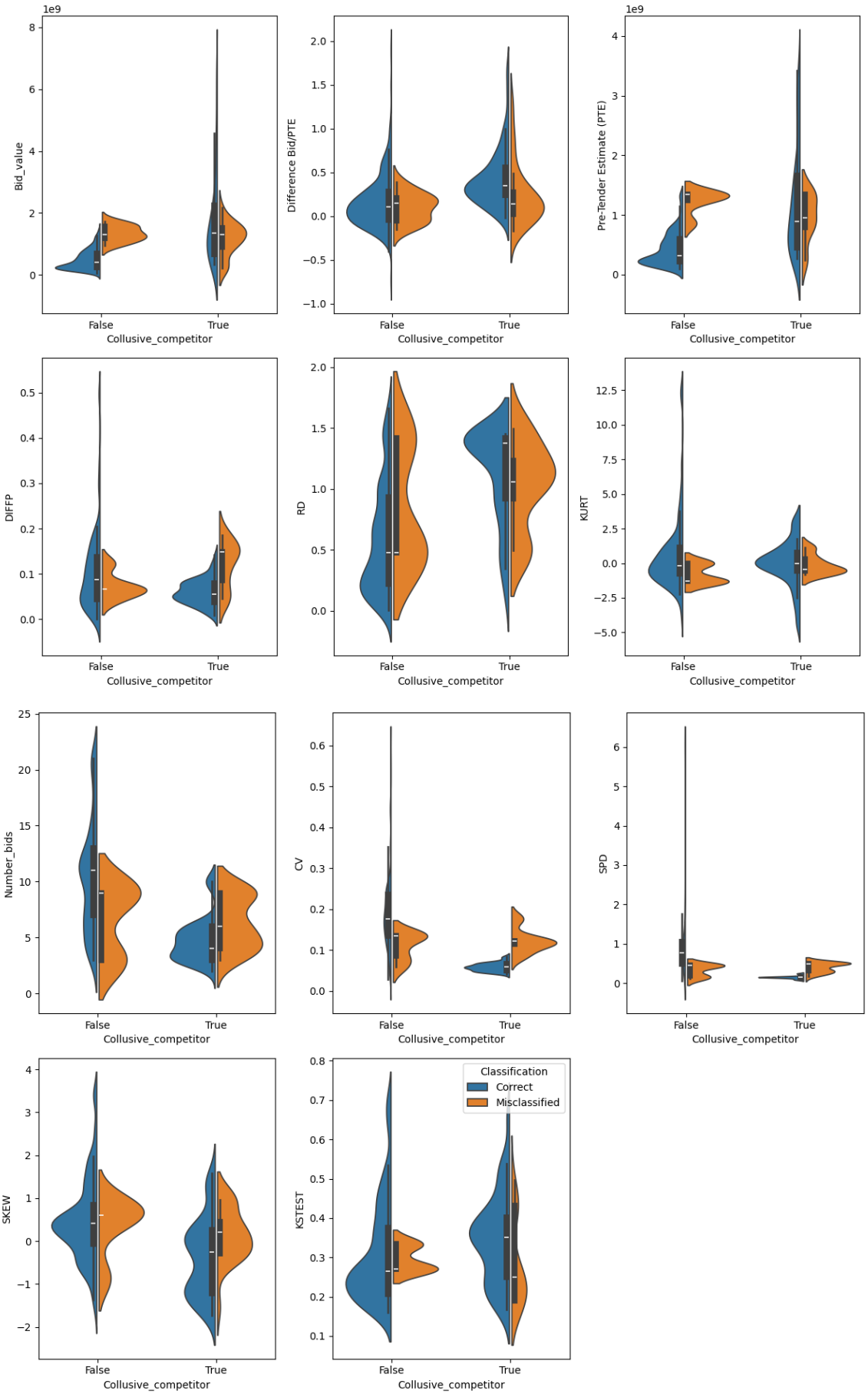
5.3 ANÁLISE DA DISTRIBUIÇÃO DOS RESULTADOS

5.3.1 Conjunto de dados brasileiro

Podemos analisar a partir do gráfico *violin plot* na Figura 10, a distribuição dos valores das *features*. A cor azul indica as distribuições de determinada *feature* classificada corretamente, enquanto que as de cor laranja indicam o contrário. Em um cenário ideal onde o modelo previsse perfeitamente, não haveria diferença entre as classificações dos valores para *False* ou *True* para a variável-alvo. Em outras palavras, quanto mais diferentes forem as distribuições azuis e laranjas, há probabilidade que o modelo cometa erros de classificação, especialmente em casos em que a importância dessa *feature* é alta.

Em uma primeira análise, verifica-se que o modelo conseguiu classificar dados discrepantes corretamente, como na *feature Number Bids*. No entanto, não pôde detectar padrões em distribuições com dois ou mais picos, como na *feature* PTE, onde a distribuição classificada incorretamente como não colusiva seguia um padrão bimodal. Similarmente, o modelo detectou padrões em uma distribuição de valores na *feature Bid Value*, mas classificou incorretamente outro pico da distribuição, onde é mais alto. Tendo isso em vista, essa dificuldade pode ter contribuído para a queda no desempenho da classificação, particularmente porque a *feature* PTE é considerada a mais importante, na sequência que a concentração da classificação de CV e SPD em *outliers*, pode ter sido o motivo pelo qual o modelo não conseguiu capturar certas nuances nos dados dos lances, impactando negativamente a eficácia geral da classificação.

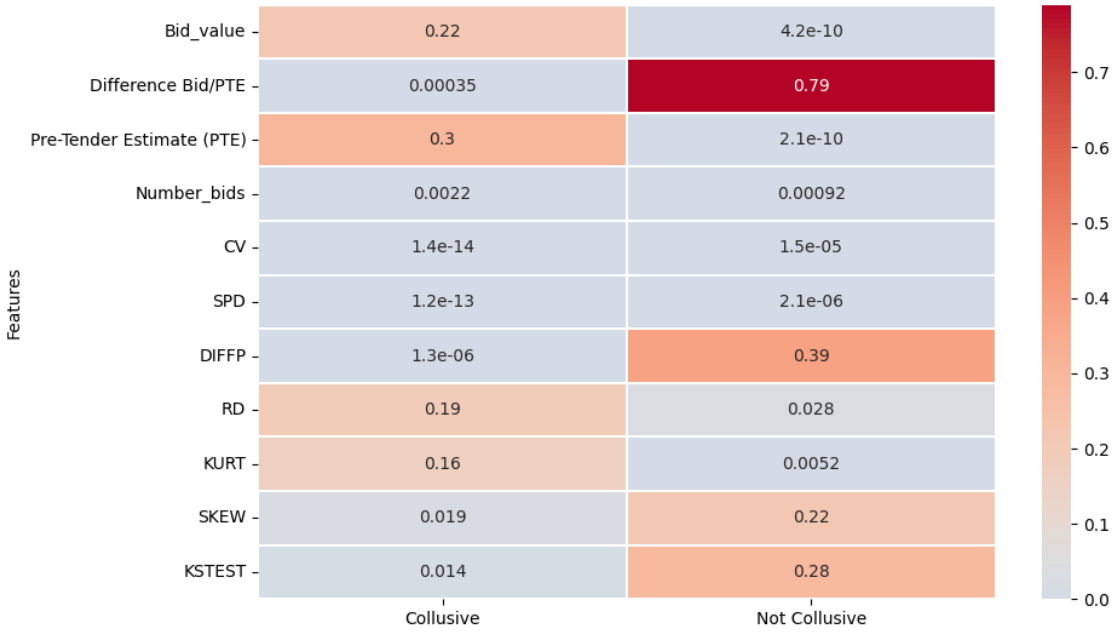
Figura 10 – *Violin plot* da distribuição da classificação das *features*



Fonte: Autor

Adicionalmente, foi realizado o teste de *Shapiro-Wilk* para avaliar se as distribuições classificadas corretamente ou não eram paramétricas, e confirmou-se que todos seguiam uma distribuição não paramétrica. Com isso, foi aplicado o teste U de *Mann-Whitney* a fim de comparar essas distribuições com base no p-value, conforme ilustrado na Figura 11. Valores de p-value acima de 0,05 estão destacados e indicam que não existem evidências para rejeitar a hipótese nula, e que as distribuições podem ser consideradas equivalentes.

Figura 11 – Teste U de Mann-Whitney da distribuição da classificação das features



Fonte: Autor

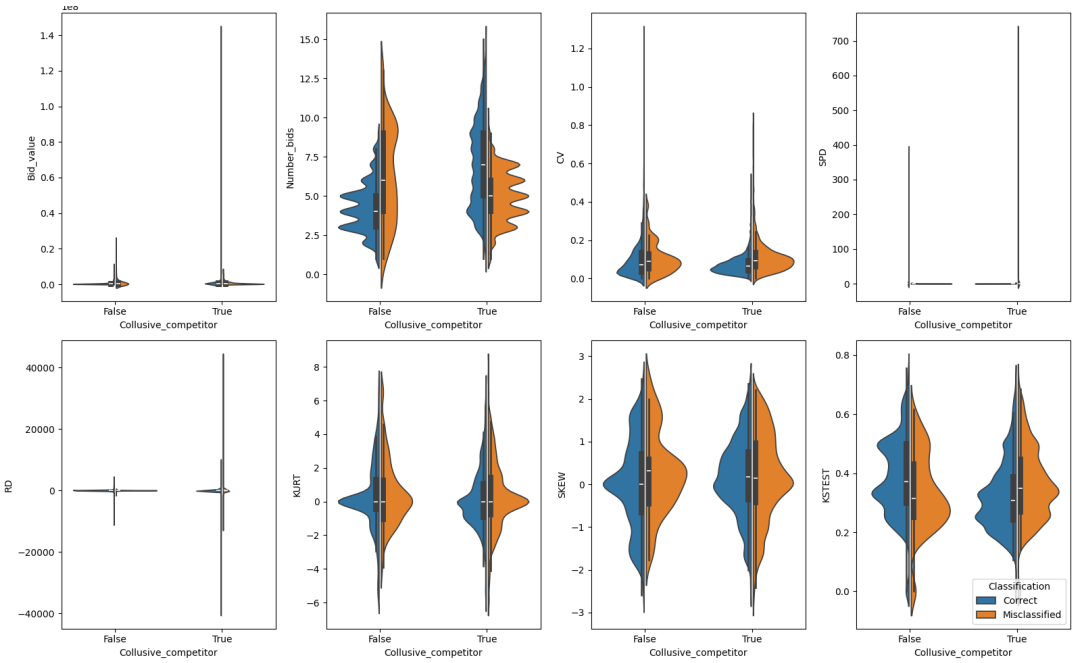
Diante desses resultados, é possível observar que os valores de p-value das distribuições das *features* mais importantes, como CV e SPD, ficaram abaixo de 0,05. Esse resultado indica não haver semelhança estatística entre as distribuições, o que costuma favorecer a classificação dos modelos, uma vez que estes podem encontrar padrões justamente na diferença das distribuições.

5.3.2 Conjunto de dados suíço

Como mencionado anteriormente, embora o modelo *Logistic Regression* pôde encontrar uma delimitação entre lances colusivos e não colusivos, observa-se na Figura 12 que o modelo se ajustou mais em valores de lances mais altos, o que pode ter prejudicado o desempenho em casos onde os lances eram menores. Outro ponto a ser destacado é a quantidade de picos nas distribuições, em particular na *feature Number Bids*, que conta com 7 picos classificados corretamente como não colusivo e 6 picos para a classificação incorreta do competidor como colusivo, o que pode sugerir que, da mesma forma, o modelo pode ter aprendido bem para

um determinado padrão na distribuição, mas no outro não. Entretanto, esse conseguiu aprender mais das nuances nas distribuições, diferentemente do *Ada Boost* no conjunto de dados brasileiro.

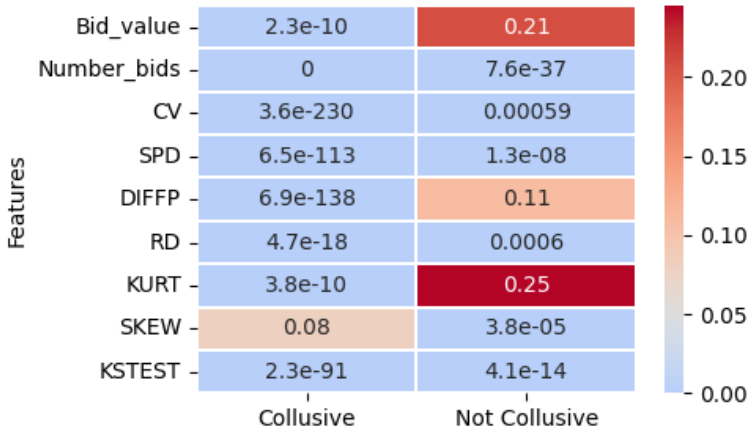
Figura 12 – *Violin plot* da distribuição da classificação das *features* (conjunto de dados suíço)



Fonte: Autor

Também pode-se observar na Figura 13 a diferença entre as distribuições a partir do teste U de *Mann-Whitney*, podendo ser apontada uma similaridade maior nas distribuições de *Bid Value*, *KURT*, *SKEW*, dentre outras. Por outro lado, *features* como *Number bids* possuem valores de p-value igual e tendendo a zero, demonstrando que as distribuições também possuem uma diferença estatisticamente significativa, que se reflete na distribuição da mesma, conforme visto anteriormente.

Figura 13 – Teste U de *Mann-Whitney* da distribuição da classificação das *features* (conjunto de dados suíço)

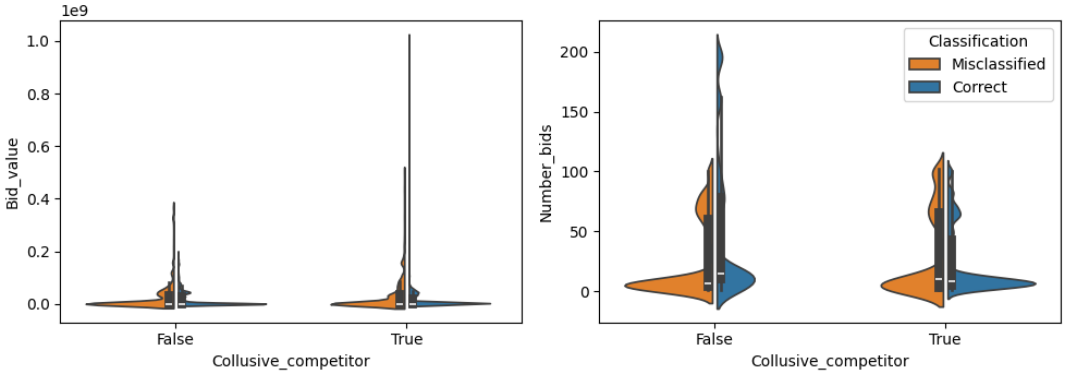


Fonte: Autor

5.3.3 Conjunto de dados agrupado

Por sua vez, para o conjunto de dados agrupado, não existe uma diferença significativa entre as distribuições para as *features* numéricas, uma vez que existem diferenças nos tipos de licitações e, por conseguinte, nos valores entre os conjuntos. Contudo, percebe-se uma distribuição bimodal para a *feature Number Bids*, havendo uma concentração maior abaixo de 50 lances, conforme ilustrado na Figura 14.

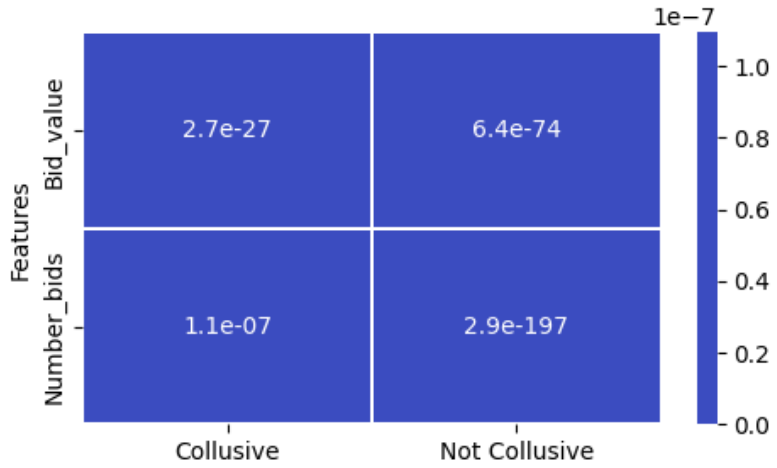
Figura 14 – *Violin plot* da distribuição da classificação das *features* (conjunto de dados agrupado)



Fonte: Autor

Ao analisar os valores de p-value do gráfico de *Mann-Whitney* para o conjunto de dados agrupado na Figura 15, nota-se que não é possível descartar a hipótese de que há diferença entre as distribuições dos lances classificados corretamente dos incorretos. Ainda que não haja assimetria entre os valores a ponto de facilitar o modelo a encontrar padrões, existe uma variabilidade acentuada, como pode ser observado nas caudas longas das distribuições, fazendo o valor de p-value ser muito baixo e, portanto, haver probabilidade de que as distribuições possuam uma grande diferença.

Figura 15 – Teste U de *Mann-Whitney* da distribuição da classificação das *features* (conjunto de dados agrupado)



Fonte: Autor

6 CONCLUSÃO

Neste trabalho, foi estudado acerca da criação de um fluxo de trabalho para auxiliar na tomada de decisão em relação aos melhores modelos de *machine learning* para ser aplicado em conjuntos de dados de conluio em licitações. Para isso, foram realizados experimentos em conjuntos de dados de conluio de diferentes países, sendo eles, Brasil, Suíça e um conjunto de diferentes países agrupados. Com isso, foram reproduzidos os resultados originais com *random subsampling*. Ao observar a variabilidade dos resultados e a falta de informações sobre o desvio padrão reportado pela literatura, foi adaptada a reprodução para serem perfeitamente replicáveis. Em seguida, foram otimizados e selecionados os modelos com maior desempenho, utilizando técnicas de validação cruzada, como *GroupKFold* ou LOGOCV. Por fim, foram analisadas a importância das *features* com base na análise da distribuição das classificações e em testes estatísticos.

Ao realizar a reprodução dos resultados com *random subsampling* para o conjunto de dados brasileiro e posteriormente com LOGOCV, foi observado um aumento no desempenho dos modelos com os hiperparâmetros já disponibilizados pelo artigo de referência, contudo, evidenciou-se um desvio padrão acentuado que, mesmo após realizar uma busca por melhores hiperparâmetros, não foi obtido um resultado capaz de reduzi-lo consideravelmente.

Também constatou-se que este fluxo de trabalho pode auxiliar não somente para a seleção do melhor modelo, como também na análise da qualidade dos dados obtidos. Por meio da análise da variabilidade dos resultados do conjunto de dados brasileiro, constatou-se que a tarefa de encontrar um melhor modelo seria mais desafiadora, sendo necessárias estratégias voltadas a melhorar o conjunto antes de otimizar modelos.

Ademais, o fluxo proposto foi aplicado em outros conjuntos de dados, auxiliando na tomada de decisão em relação à qualidade dos dados e modelos, como demonstrado para conjunto de dados suíço e agrupado. Podemos concluir a partir de ambos que ter uma maior quantidade de licitações e lances resulta em um desvio padrão menor. Além disso, o uso de modelos mais simples também pode apresentar um desempenho e desvio padrão satisfatórios. Entretanto, é preciso avaliar se as distribuições das classificações feitas pelo modelo estão extraindo padrões em diferentes faixas de valores.

6.1 TRABALHOS FUTUROS

Embora este fluxo tenha se proposto a melhorar os resultados dos modelos, concluiu-se que existe a necessidade de aprimorar inicialmente o conjunto de dados. Dessa forma, pode-se considerar a elaboração de trabalhos no intuito de enriquecer os conjuntos utilizados neste estudo para elevar o desempenho das predições. Em combinação com uma quantidade maior de dados, o uso de técnicas que visem um balanceamento maior da variável alvo pode ser importante para o desempenho do modelo, como Synthetic Minority Oversampling Technique (SMOTE).

Além disso, o uso de técnicas concentradas em adaptar o problema para Redes Neurais pode ser promissora para este contexto, haja vista que, embora não tendo o melhor resultado, o modelo MLP foi o que obteve o maior ganho de desempenho e redução no desvio padrão.

REFERÊNCIAS

ABRANTES-METZ, R. M. Roundtable on ex officio cartel investigations and the use of screens to detect cartels. 2013.

ANOWAR, F.; SADAOUI, S. Multi-class ensemble learning of imbalanced bidding fraud data. p. 352–358, 2019.

BRASIL. Lei nº 14.133/2021. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 2021. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm.

FAZEKAS, M.; TóTH, I. J.; KING, L. P. Corruption manual for beginners:'corruption techniques' in public procurement with examples from hungary. **Corruption Research Center Budapest Working Paper no. CRCB-WP/2013**, v. 1, 2013.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine learning**, Springer, v. 63, p. 3–42, 2006.

GUIMARãES, E. dos S. **Manual de planejamento das licitações públicas**. Curitiba, PR: Juruá, 2012.

HENDRICKS, K.; PORTER, R. H. Collusion in auctions. **Annales d'Economie et de Statistique**, JSTOR, p. 217–230, 1989.

JUNIOR, J. T. P. **Critérios de julgamento**. 2022. Disponível em: https://www.novaleilicitacao.com.br/2020/01/20/criterios-de-julgamento/.

MARGINEANTU, D. D.; DIETTERICH, T. G. Pruning adaptive boosting. In: CITESEER. **ICML**. [S.l.], 1997. v. 97, p. 211–218.

MITCHELL, T. M. **Machine learning**. 1997.

MORO, S. F. Preventing systemic corruption in brazil. **Daedalus**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 147, n. 3, p. 157–168, 2018.

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. **Frontiers in neurorobotics**, Frontiers Media SA, v. 7, p. 21, 2013.

OECD. **Collusion: Competition Policy in the Digital Age**. OECD: Paris, France, 2017. Disponível em: https://www.oecd.org/daf/competition/Algorithms-and-colllusion-competition-policy-in-the-digital-age.pdf.

PENA, F. L. Especialização em Gestão Pública. **PLANEJAMENTO DAS LICITAÇÕES: Um Estudo de Caso em uma Empresa Pública**. Sete Lagoas, MG: [s.n.], 2019. Disponível em: https://repositorio.ufmg.br/bitstream/1843/32320/1/TCC%20FELIPE%20LOPES%20PENA%20-%20GP.pdf.

PIETRO, M. S. Z. D. **Direito Administrativo**. São Paulo, SP: Editora Forense, 2016.

PORTER, R. H.; ZONA, J. D. Detection of bid rigging in procurement auctions. **Journal of political economy**, The University of Chicago Press, v. 101, n. 3, p. 518–538, 1993.

RABUZIN, K.; MODRUSAN, N. Prediction of public procurement corruption indices using machine learning methods. In: **KMIS**. [S.l.: s.n.], 2019. p. 333–340.

RIEDMILLER, M.; LERNEN, A. Multi layer perceptron. **Machine Learning Lab Special Lecture, University of Freiburg**, v. 24, 2014.

RODRíGUEZ, M. J. G. et al. Collusion detection in public procurement auctions with machine learning algorithms. **Automation in Construction**, Elsevier, v. 133, p. 104047, 2022. Disponível em: https://www.sciencedirect.com/science/article/pii/S0926580521004982.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. [S.l.]: Pearson Education India, 2016.

THORSTENSEN, L. F. G. V. **Brasil na OCDE: Compras Públicas**. 2021.

VELASCO, R. B. et al. A decision support system for fraud detection in public procurement. **International Transactions in Operational Research**, Wiley Online Library, v. 28, n. 1, p. 27–47, 2021.

WALLIMANN, H.; IMHOF, D.; HUBER, M. A machine learning approach for flagging incomplete bid-rigging cartels. **Computational Economics**, Springer, p. 1–52, 2022.

WONG, T.-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. **Pattern Recognition**, v. 48, n. 9, p. 2839–2846, 2015. ISSN 0031-3203. Disponível em: https://www.sciencedirect.com/science/article/pii/S0031320315000989.

ZHOU, Z.-H. **Machine learning**. [S.l.]: Springer Nature, 2021.

APÊNDICE A – INSTRUÇÕES PARA EXECUÇÃO DO FLUXO DE TRABALHO

O código para a reprodução do fluxo proposto neste trabalho pode ser encontrado no link https://github.com/LeonardoVieiraNunes/collusion-detection-workflow.

Para a correta execução, é necessário ter a versão 3.10 do *Python*, o gerenciador de dependências *pip* e o *Jupyter Notebook*. Após instalado, execute o seguinte comando para instalar as dependências necessárias:

```
pip install -r requirements.txt
```

Com o Jupyter Notebook aberto e devidamente configurado com a versão do *Python*, abra o arquivo *optimization* ou *data analysis* e clique no botão para executar todas as células.

O arquivo *optimization* contém os espaços de hiperparâmetros, bem como as regras para embaralhamento e seleção de *features* a depender do conjunto de dados, aos quais estão armazenados no diretório *datasets*. Após a reprodução, o melhor conjunto de hiperparâmetros para determinado modelo e conjunto de dados são armazenados no diretório *results* correspondente ao nome do conjunto de dados.

Por sua vez, o arquivo *data analysis* possui todas as etapas de análises dos melhores modelos. Para selecionar o modelo específico, basta inserir os modelos e parametros desejados na variável *modelos* na célula 14.

Para uso deste fluxo em outros conjunto de dados, é necessário realizar alguns ajustes no código para adequar a variável alvo e as demais *features*, especialmente na função *cast features* no arquivo *optimization*

A.1 UTILITÁRIOS

No diretório *utils* poderão ser encontrados os arquivos utilitários para a criação dos gráficos neste projeto, como os gráficos de comparação do *random subsampling* reportado pelos autores e o reproduzido neste trabalho, assim como o gráfico comparativo dos resultados com os hiperparâmetros originais e os encontrados utilizando o fluxo proposto.

APÊNDICE B – ARTIGO SBC

Detecção de conluio em licitações utilizando algoritmos de machine learning

Leonardo Vieira Nunes[1]

[1]Departamento de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Florianópolis – SC – Brazil

leonardovnun@gmail.com

Abstract. *The collusion consists of a criminal practice where companies secretly agree on bid values among themselves in a particular tender with the aim of maximizing their profit. This type of practice inflates prices and harms the quality of products and services acquired by the public administration, directly impacting the lives of citizens. In this context, approaches using machine learning methods have emerged that analyze large amounts of data to detect the presence of collusion. However, the lack of a pre-established workflow to analyze and compare the behavior of algorithms and datasets makes decision-making more challenging. Therefore, this study aims to implement a workflow to assist in decision-making regarding the best machine learning algorithms to be used for a specific dataset. To achieve this, results found in the literature were reproduced, the hyperparameters of the models were optimized, the one with the best balanced accuracy was selected, and finally, the importance of the features, distribution of classifications by features, and Mann-Whitney U Test were analyzed. From these results, it was concluded that there is a need for enrichment in smaller datasets, as it resulted in a high standard deviation of balanced accuracy. It was also noted that there was a significant performance gain in models trained with larger datasets, with a balanced accuracy of around 83% for St. Gallen and Graubünden.*

Resumo. *O conluio consiste em uma prática criminosa onde empresas concordam secretamente valores de lances entre si em determinada licitação com o objetivo de maximizarem o seu lucro. Esse tipo de prática inflaciona os preços e prejudica a qualidade dos produtos e serviços adquiridos pela administração pública, impactando diretamente na vida dos cidadãos. Nesse contexto, surgiram abordagens utilizando métodos baseados em aprendizado de máquina que analisam grandes quantidades de dados para detectarem a presença de conluio. Todavia, a falta de um fluxo pré-estabelecido para analisar e comparar o comportamento dos algoritmos e conjuntos de dados torna a tomada de decisão mais desafiadora. Diante disso, este estudo busca implementar um fluxo de trabalho para auxiliar na tomada de decisão em relação aos melhores algoritmos de machine learning a serem utilizados para determinado conjunto de dados. Para isso, foram reproduzidos resultados encontrados na literatura, otimizados os hiperparâmetros dos modelos, selecionado o que obteve a melhor acurácia balanceada e, por fim, analisado a importância das features, distribuição das classificações pelas features e o teste U de Mann-Whitney. A partir desses resultados, concluiu-se que há a necessidade de enriquecimento em conjuntos*

menores, uma vez que resultou em um alto desvio padrão da acurácia balanceada. Também notou-se que houve um ganho de desempenho significativo nos modelos treinados com conjuntos maiores, apresentando uma acurácia balanceada em torno de 83% para o de St. Gallen e Graubünden.

1. Introdução

O processo de licitação consiste em uma ferramenta utilizada globalmente pelas administrações públicas para adquirir bens e serviços do setor privado. De acordo com [Vera Thorstensen 2021], estima-se que no Brasil, cerca de 12% do Produto Interno Bruto (PIB) tenha sido destinado a compras públicas entre os anos de 2002 a 2019, sendo a sua maioria por meio de processos licitatórios, o que demonstra ser um volume significativo de recursos públicos que passam por esse processo.

Embora existam leis destinadas a promover a livre concorrência entre os participantes, a prática de fraudes em licitações é uma ocorrência frequente e complexa de ser detectada pelos órgãos de controle. Essas fraudes podem assumir formas diretas, com empresas conspirando para fixar preços, ou indiretas, especialmente em mercados dominados por oligopólios, nos quais poucas empresas têm a capacidade de participar de licitações específicas. Isso resulta em uma série de problemas, como obras e serviços públicos de qualidade inferior para a população, redução da concorrência, dentre outros. [OECD 2017].

Devido à abundância de dados gerados por esses processos nos últimos anos em forma digital, surgiram soluções que obtiveram êxito ao utilizar métodos avançados de análises de dados, como *machine learning*, para auxiliar na identificação de padrões fraudulentos. Dentre eles, está o trabalho de [Velasco et al. 2021] onde foi implementado um sistema de suporte à decisão baseado em *machine learning* o qual auxiliou na investigação de contratos supostamente fraudulentos avaliados em cerca de R$ 3,6 bilhões.

Ainda que tais métodos tenham aperfeiçoado a identificação de fraudes, ainda existem desafios na escolha e ajuste de modelos para determinado caso. [Anowar and Sadaoui 2019] elencam problemas como desbalanceamento de classes, similaridade de lances fraudulentos com comportamentos não fraudulentos e otimização de desempenho, que podem dificultar a análise do desempenho de modelos preditores.

Nesse sentido, este trabalho se propõe a analisar as atuais técnicas de detecção de fraudes em licitações utilizando dados disponibilizados publicamente a fim de auxiliar na decisão de quais modelos podem ser utilizados para determinado conjunto de dados, assim como apontar a necessidade de aperfeiçoá-lo. Para isso, serão utilizadas técnicas de otimização em modelos de *machine learning*, bem como a validação dos resultados das classificações realizadas por eles em diferentes conjuntos de dados.

2. Objetivos

2.1. Objetivo geral

Analisar a eficácia de modelos de *machine learning* para detectar fraudes em licitações públicas e criar um fluxo de trabalho que permita aprimorar o desempenho e facilitar a tomada de decisão para a escolha de qual modelo utilizar em ambientes de produção.

2.2. Objetivos específicos

1. Realizar revisão bibliográfica de técnicas de detecção de fraudes;
2. Reproduzir os experimentos encontrados na literatura;
3. Selecionar modelos com maior desempenho e realizar otimizações em cada um deles;
4. Avaliar o resultado do desempenho dos modelos otimizados em relação à literatura;
5. Analisar a variabilidade dos resultados usando diferentes estratégias de divisão dos dados;
6. Analisar a importância das *features* do melhor modelo encontrado.

3. Trabalhos relacionados

Com o avanço no poder de processamento de dados e *data mining*, [Velasco et al. 2021] propuseram no artigo *"A decision support system for fraud detection in public procurement"* um sistema para auxiliar órgãos de monitoramento com indícios de fraudes em licitações. Para isso, os autores elaboraram um conjunto de comportamento que poderiam indicar uma prática fraudulenta na licitação, denominado *risk patterns*.

Neste estudo, os autores apresentam casos operacionais dessa metodologia, uma delas em que foram analisadas 857 empresas que venceram contratos de licitação, mas que possuíam empresas em comum disputando o mesmo edital, somando uma quantia de mais de R$ 3,6 bilhões em contratos. Contudo, os autores reconhecem a importância de uma análise manual para interpretar os resultados do modelo, bem como destacam suas limitações, como a a necessidade de dados abundantes e de alta qualidade.

Por sua vez, no artigo *"Prediction of public procurement corruption indices using machine learning methods"*, [Rabuzin and Modrusan 2019] aplicam técnicas de *data mining* para extrair informações de licitações em órgãos públicos da Croácia. Utilizando um modelo baseado em processamento de linguagem natural (MLP) para extrair variáveis como condições técnicas, prazos e valores estimados. Essas variáveis são então utilizadas para treinar modelos de *machine learning* como *support vector machine*, *naive bayes* e regressão logística, visando descobrir padrões nos dados automaticamente.

A avaliação dos modelos inclui métricas como acurácia, precisão, *recall* e ROC, demonstrando resultados de precisão variando entre 60% e 85%, com *recall* acima de 5%, considerado satisfatório pelos autores. No entanto, o estudo enfrentou limitações como o tamanho pequeno do conjunto de dados disponível e a falta de preditores específicos, levando os autores a sugerir o uso de modelos mais avançados como redes neurais e aprendizado profundo para melhorar os resultados.

Devido ao surgimento de soluções utilizando *machine learning* para detectar fraudes em licitações, [Rodríguez et al. 2022] realizaram um estudo transversal para testar o desempenho de diferentes combinações de modelos, conjuntos de dados e configuração de *features*. Utilizando dados de contratos de licitação de vários países, incluindo casos da operação Lava Jato no Brasil em 2014, os autores treinaram e validaram 11 modelos distintos.

Os resultados mostraram um desempenho geralmente acima de 80% de acurácia balanceada, índices de falsos positivos e falsos negativos abaixo de 10%. No entanto,

os autores identificaram limitações como a necessidade de otimização mais detalhada dos hiperparâmetros, inclusão de mais variáveis e dados para aprimorar ainda mais os modelos.

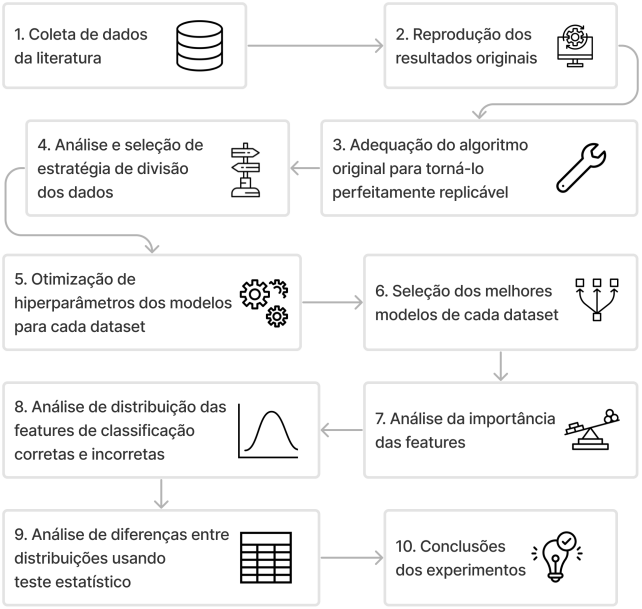
Por fim, no artigo *"A machine learning approach for flagging incomplete bid rigging cartels"*, [Wallimann et al. 2022] realizaram o uso de técnicas de regressão utilizando *machine learning* em conjuntos de dados de fraudes em licitações nas regiões de See-Gaster e Graubünden, na Suíça, a fim de estimar a probabilidade de cartéis estarem participando de licitações através dos lances onde nem todos os concorrentes da licitação estavam envolvidos no esquema.

Após a validação, foram comparadas as taxas de classificação correta em relação a um modelo de referência para diferentes amostras e configurações das *features*, a qual variou entre 61,2% e 84,1%. Os autores também ressaltaram que esta diferença se deu por conta de supostos casos onde o cartel eventualmente não praticava o conluio total em determinada licitação. Além disso, o uso de configurações de amostras onde havia o valor do lance e a quantidade de licitantes como preditores aumentaram as taxas de classificação na ordem de 5 a 10 pontos percentuais em relação aos demais.

4. Metodologia

A metodologia deste trabalho consiste em realizar análises sobre o conjunto de dados disponibilizado no estudo de [Rodríguez et al. 2022], aplicar otimizações sobre os modelos com melhor desempenho evidenciado pelos autores, além de testar modelos adicionais e, por fim, analisar os resultados obtidos. A motivação para a utilização deste conjunto em específico se dá pela riqueza de informações sobre diferentes casos de fraudes em licitações em diversos países. Além disso, trata-se de casos de fraudes recentes e com informações amplamente divulgadas, facilitando a compreensão dos dados. A Figura 1 ilustra o fluxo aplicado neste e em outros dois conjuntos de dados desse mesmo artigo de referência.

Figure 1. Fluxograma da metodologia realizada neste trabalho



4.1. Coleta e Detalhamento dos conjuntos de dados

Nesta primeira etapa do fluxo de trabalho, os conjuntos de dados utilizados e o código original foram obtidos através do material suplementar fornecido no artigo de [Rodríguez et al. 2022]. Sendo essa a primeira etapa no fluxo de trabalho, apresentado na Figura 1. Dentre eles, o conjunto de conluio em licitações de obras da empresa Petróleo Brasileiro (Petrobras) foi o objeto do estudo mais aprofundado. Além deste, foram realizados experimentos adicionais com o conjunto de dados suíço, que contém um número consideravelmente maior de registros. E por fim, um experimento com todos os conjuntos do trabalho de referência agrupados.

O conjunto de dados brasileiro representa o resultados da investigação da Polícia Federal em 2014, onde foi apontado um esquema de lavagem de dinheiro por meio de contratos de licitações da Petrobras. Neste conjunto de dados, existem 683 lances com 272 licitantes, distribuídos em 101 licitações de contratos de obras e serviços.

Já o conjunto de dados suíço representa os dados obtidos de um esquema de conluio em licitações de infraestrutura rodoviária nos cantões suíços de St. Gallen e Graubünden (SG&GR) entre os anos de 2004 a 2010. Esse conjunto de dados possui 4.344 licitações e 21.231 lances. A motivação para a escolha deste se deu pela quantidade e qualidade dos dados, uma vez que, diferentemente do conjunto de dados agrupado, os dados são estruturados para apenas um escopo de fraude, além de contar com as seguintes *features*: *Tender*, *Bid Value*, *Winner*, *Number Bids*, *Date*, *Contract type* e *Collusive competitor*.

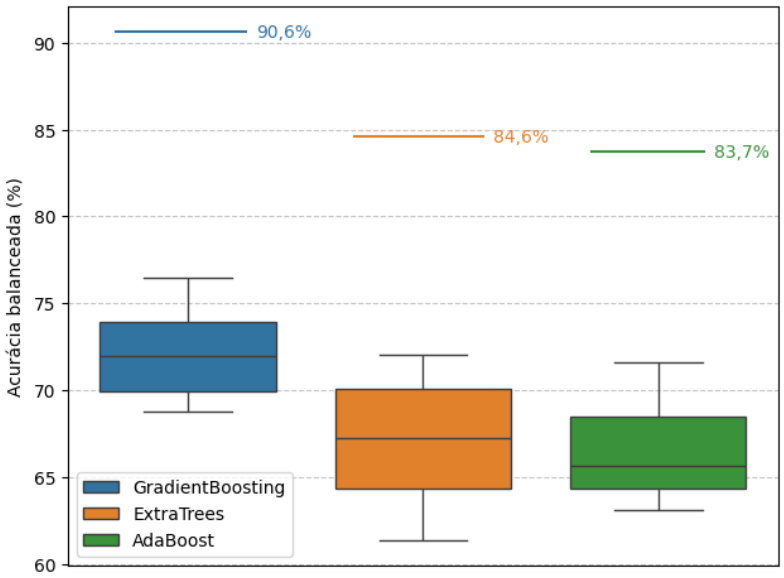
Outro conjunto de dados objeto desse estudo foi o que reúne todos os outros conjuntos estudados pelo artigo de referência em apenas um. Esse conjunto possui 9.781 licitações e 64.348 lances compilados em seis conjuntos diferentes. A escolha deste conjunto se deu pela quantidade de licitações e lances, algo que o conjunto de dados brasileiro carece. A quantidade de *features* utilizadas foi limitada àquelas utilizadas pelo autor, a fim de ser possível comparar os resultados, sendo elas *Tender*, *Bid Value*, *Winner*, *Number Bids* e *Dataset*. Esta última *feature* corresponde ao identificador do conjunto de dados de onde foi extraído o dado.

4.2. Reprodução dos resultados

Nesta segunda etapa do fluxo de trabalho, foi constatado que, no código original, os resultados foram obtidos por meio de 50 iterações de treinamento e teste dos modelos. O desempenho consistiu na média dessas iterações, técnica esta chamada de *random subsampling* [Tan et al. 2016]. Todavia, não foram reportados os desvios padrão, e nem fixadas, no código disponibilizado, as sementes para geração de números aleatórios, de forma que a reprodução exata dos resultados pudesse ser realizada.

A Figura 2 apresenta a acurácia balanceada obtida neste estudo após 10 execuções do algoritmo original para os três melhores modelos. As linhas com as respectivas cores nas legendas correspondem aos resultados reportados pelos autores. Como pode ser observado, a acurácia balanceada do artigo é na ordem de 10 a 20% acima da encontrada após a reprodução dos resultados neste estudo. Em teoria, a tendência geral dos resultados deveria se manter mesmo com diferentes *seeds*. No entanto, devido à ausência de informações sobre o desvio padrão no artigo de referência, não haveria como confirmar se essa variabilidade seria aceitável no desvio padrão obtido pelos autores.

Figure 2. Boxplot da reprodução dos resultados originais



Devido à grande variabilidade dos resultados e a quantidade limitada de dados, foi utilizada a técnica *Leave One Group Out Cross Validation* (LOGOCV), que é uma forma especial de validação cruzada semelhante ao método *k-fold*. Neste contexto, cada iteração usa todas as licitações, exceto uma, para treinamento, e a licitação excluída para teste, evitando contaminação entre conjuntos de treinamento e teste. De acordo com [Wong 2015], essa técnica é comumente aplicada em conjuntos de dados pequenos para obter resultados consistentes sem variabilidade na divisão dos dados.

4.3. Otimização de hiperparâmetros dos modelos

Na quinta etapa, foram otimizados os três melhores modelos do artigo de referência, além dos modelos *Multi Layer Perceptron* (MLP) e *Logistic Regression*, para comparar os resultados. O MLP, embora não tenha obtido os melhores resultados no artigo original, foi incluído por sua ampla utilização em estudos de classificação. Já o *Logistic Regression* foi escolhido por ser um modelo fundamental em técnicas de classificação, servindo como critério para estabelecer um limite inferior dos resultados dos demais modelos.

4.4. Análise dos resultados

Na sétima etapa do fluxo de trabalho, após a seleção dos melhores modelos, foram utilizados métodos para analisar os resultados, de modo que fosse possível extrair conhecimento adicional do processo de treinamento e teste dos modelos, evidenciando as características mais importantes utilizadas por eles para a detecção de conluio. Para isso, foi extraída a *feature importance* dos melhores modelos de cada conjunto de dados. Essa métrica compreende a importância média que cada *feature* teve na utilização para a realização da detecção por parte dos modelos. A interpretação destes valores por especialistas do domínio do problema, isto é, promotores, auditores, fiscais e etc., tem especial valor para aprimorar a prevenção e detecção de fraudes em licitações públicas.

Além disso, de forma a melhor entender os erros de classificação dos melhores modelos, as licitações foram divididas entre as classificadas correta e incorretamente,

e posteriormente foram analisadas as distribuições dos valores de cada *feature* dessas classificações. Para isso, tais distribuições foram ilustradas em gráficos *violin plots* para cada *feature* numérica utilizada nos modelos.

A partir da distribuição das classificações corretas e incorretas de cada *feature*, foram realizados testes estatísticos para identificar a influência de cada *feature* na classificação final. Esse tipo de análise tem por objetivo verificar, com um intervalo de confiança de 95%, se existe diferença entre as duas distribuições. Para isso, todas as amostras foram inicialmente submetidas ao teste de *Shapiro-Wilk*, cujo objetivo é comparar se essas distribuições seguem uma distribuição normal. Caso afirmativo, são analisadas a partir do teste t de *Student*. Do contrário, são comparadas a partir do teste U de *Mann-Whitney*.

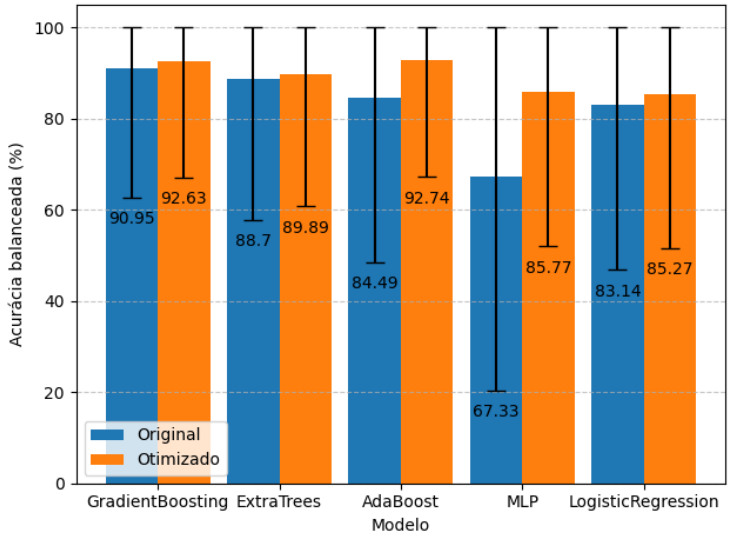
5. Resultados

Nesta seção apresentamos os resultados experimentais obtidos ao aplicar o fluxo de trabalho proposto na metodologia para analisar os modelos treinados utilizando o conjunto de dados brasileiro com LOGOCV e os outros dois conjuntos com *GroupKFold* com 50 *folds*.

5.1. Otimização de hiperparâmetros

A Figura 3 apresenta os resultados da aplicação do LOGOCV usando os mesmos hiperparâmetros do artigo de referência e comparando com os resultados da otimização deste estudo. Observa-se um ganho geral de desempenho em todos os modelos, sendo o *Ada Boost* o melhor, com 92,74%, ficando próximo ao resultado do *Gradient Boosting*, com 92,63%. Além disso, o MLP foi o que obteve o maior ganho de desempenho, aumentando sua acurácia balanceada em 18,44%. Todavia, nota-se que o desvio padrão ainda continua alto, o que demonstra que mesmo com a otimização dos resultados, os modelos ainda não possuem uma constância na predição das licitações.

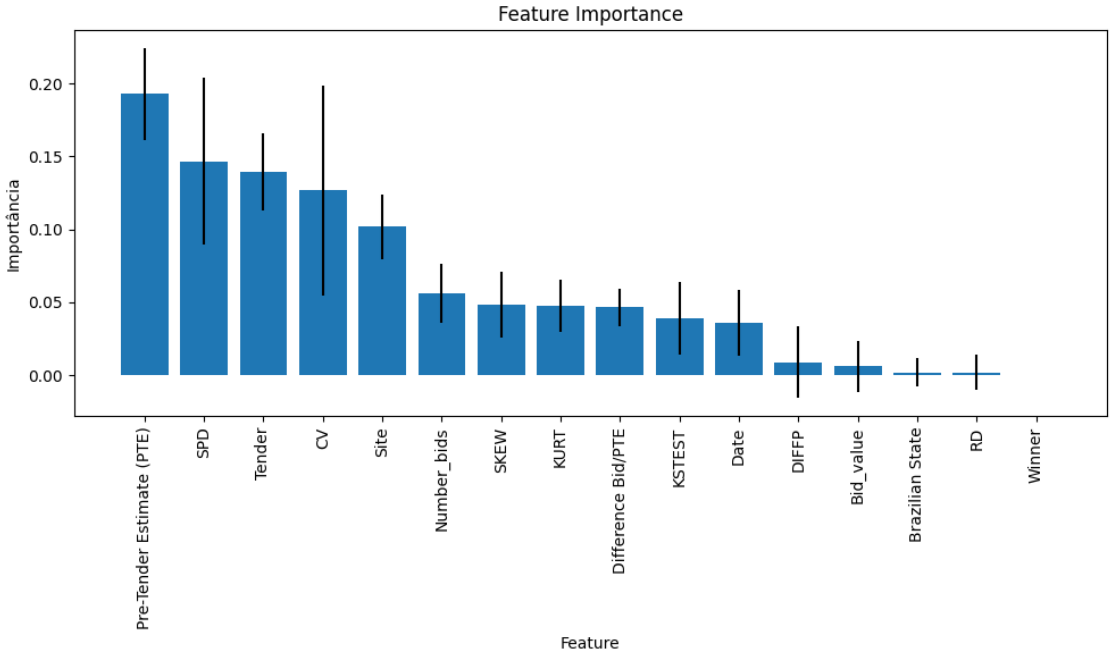
Figure 3. Otimização dos hiperparâmetros com LOGOCV



5.2. *Feature importance*

Após selecionado o melhor modelo, foram realizadas algumas análises a fim de compreender os resultados obtidos, dentre elas está a importância média das *features*. A Figura 4 representa a *feature importance* do *Ada Boost*.

Figure 4. *Feature importance* (*Ada Boost*)



Examinando a figura, podemos observar que o modelo utilizou *features* de diferentes categorias, sem concentrar a importância em um conjunto reduzido. As mais relevantes foram PTE, SPD, *Tender*, CV e *Site*. O PTE sugere que o potencial de fraudes é maior em licitações com lances mais altos, indicando que empresas colusivas focam mais nessa característica de licitação. SPD e CV com maior importância sugere que a variabilidade nos lances pode indicar comportamentos colusivos, como acordos prévios para oferecer lances com valores mais baixos. No caso de *Tender* e *Site* indicam que pode haver localizações com maior frequência de conluio entre empresas.

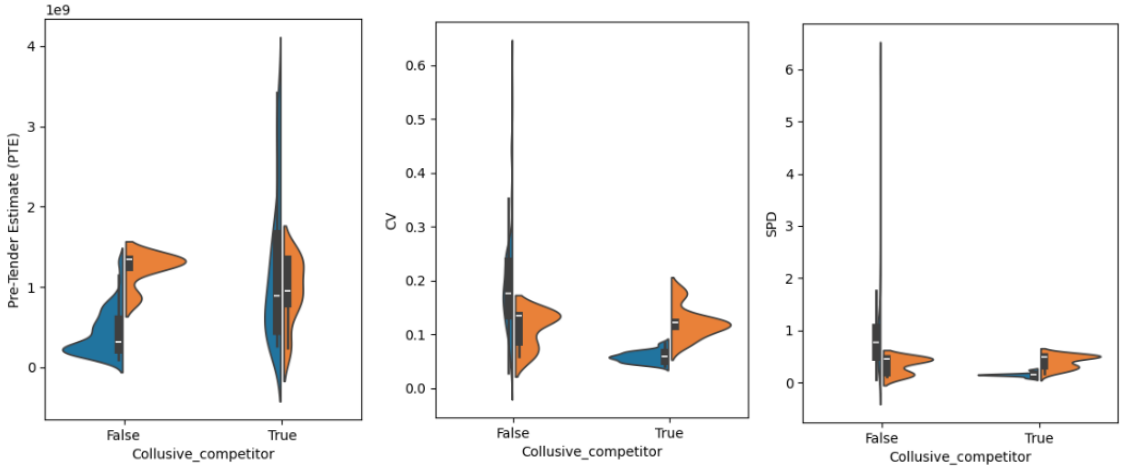
5.3. Análise da distribuição dos resultados

Neste passo, os dados foram separados entre os classificados corretamente dos incorretos, de modo a elaborar um gráfico de *violin plot* das features, conforme pode ser observado na Figura 5, onde em azul estão as classificações corretas e em laranja as incorretas. Quanto mais diferentes forem as distribuições, maior a probabilidade que o modelo cometa erros de classificação, especialmente em casos em que a importância dessa *feature* é alta.

Podemos observar que o modelo teve dificuldades na classificação de distribuições com dois ou mais picos, como na *feature* PTE, onde a distribuição classificada incorretamente como não colusiva seguia um padrão bimodal. Tendo isso em vista, essa dificuldade pode ter contribuído para a queda no desempenho da classificação, particularmente porque a *feature* PTE é considerada a mais importante, na sequência que a concentração

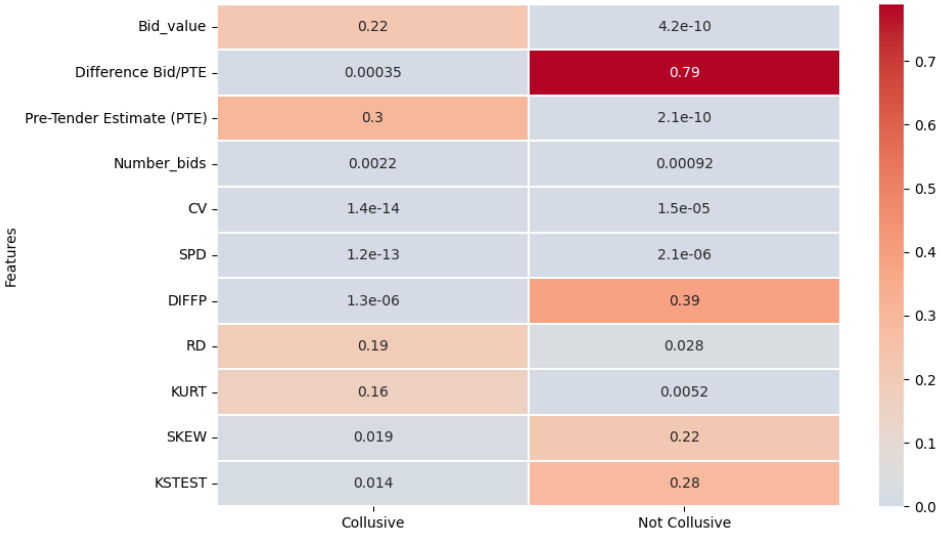
da classificação de CV e SPD em *outliers*, pode ter sido o motivo pelo qual o modelo não conseguiu capturar certas nuances nos dados dos lances, impactando negativamente a eficácia geral da classificação.

Figure 5. *Violin plot* da distribuição da classificação das *features*



Adicionalmente, foi realizado o teste de *Shapiro-Wilk* para avaliar se as distribuições classificadas corretamente ou não eram paramétricas, e confirmou-se que todos seguiam uma distribuição não paramétrica. Com isso, foi aplicado o teste U de *Mann-Whitney* a fim de comparar essas distribuições com base no p-value, conforme ilustrado na Figura 6. Valores de p-value acima de 0,05 estão destacados e indicam que não existem evidências para rejeitar a hipótese nula, e que as distribuições podem ser consideradas equivalentes.

Figure 6. Teste U de Mann-Whitney da distribuição da classificação das features



Diante desses resultados, é possível observar que os valores de p-value das distribuições das *features* mais importantes, como CV e SPD, ficaram abaixo de 0,05. Esse resultado indica não haver semelhança estatística entre as distribuições, o que cos-

tuma favorecer a classificação dos modelos, uma vez que estes podem encontrar padrões justamente na diferença das distribuições.

5.4. Resultados de outros conjuntos de dados

Nesta etapa, apresentaremos os resultados obtidos com o mesmo fluxo de trabalho aplicado no conjunto de dados brasileiro nos conjuntos de dados agrupado e suíço. Ambos utilizaram a estratégia de validação cruzada *Group K Fold* com 50 *folds*

5.4.1. Otimização de hiperparâmetros

De acordo com a Figura 7, o *Logistic Regression* foi o melhor modelo para o conjunto de dados suíço, com 83,54%. Apesar de ser o modelo mais simples, ele teve o menor desvio padrão e o maior ganho de desempenho. Isso pode ser explicado pelo fato das *features* mais importantes serem mais facilmente separadas por uma função logística. Além disso, por ser um modelo rápido de treinar, foi possível explorar mais hiperparâmetros em menos tempo, contribuindo para uma melhor adequação ao conjunto de dados.

Para o conjunto de dados agrupado, a Figura 8 mostra um pequeno ganho de desempenho, com o *Gradient Boosting* sendo o melhor, com 84,74%. Similar ao conjunto de dados brasileiro, os resultados do *Logistic Regression* estabilizaram, indicando que pode haver características não lineares nos dados que o *Gradient Boosting* explora de forma mais eficiente.

Figure 7. Otimização dos modelos (conjunto de dados suíço)

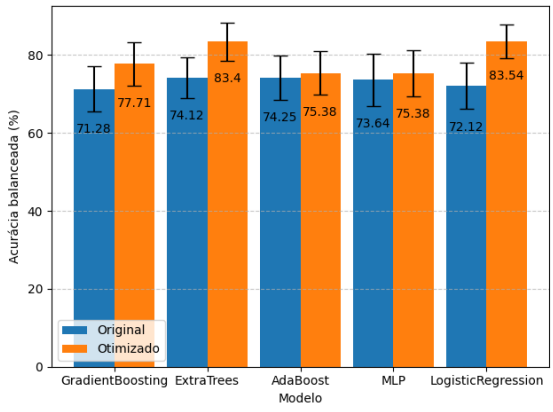
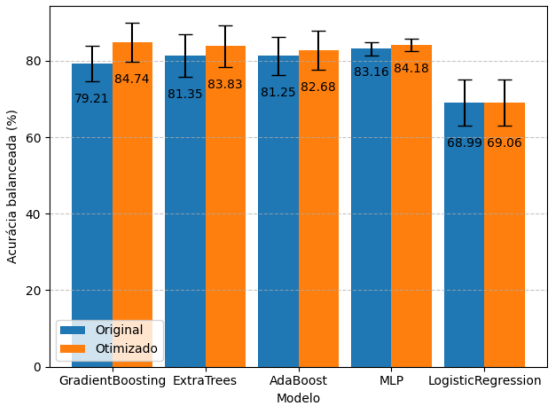


Figure 8. Otimização dos modelos (conjunto de dados agrupado)



5.4.2. *Feature importance*

Com os melhores modelos de cada conjunto de dados definidos, foi realizada a análise da importância das features de cada um, conforme representado nas Figuras 9 e 10.

Na primeira figura, a *feature Bid Value* teve o maior peso na criação da curva de decisão do modelo *Logistic Regression*, com 23% de importância, seguida pela *feature*

Number Bids, com 5%. Isso indica que o modelo conseguiu diferenciar lances colusivos dos não colusivos, atribuindo pesos significativos a essas *features*. No entanto, essa relação também afetou a taxa de erro do modelo, conforme veremos nas distribuições dos resultados.

No conjunto de dados agrupado, o modelo *Gradient Boosting* utilizou a *feature* categórica *Dataset* com 57% de importância, seguida pelas *features Number Bids* e *Bid Value*, com 23% e 16% de importância, respectivamente. Isso sugere que tanto os valores dos lances quanto a quantidade de licitantes têm uma importância significativa em ambos os conjuntos de dados, indicando a presença de conluio em determinadas licitações.

Figure 9. *Feature importance* (conjunto de dados suíço)

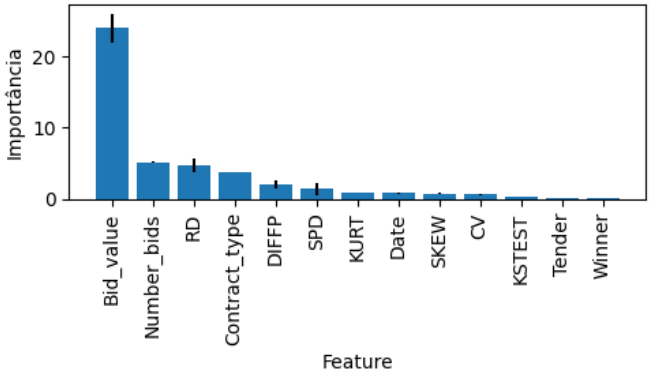
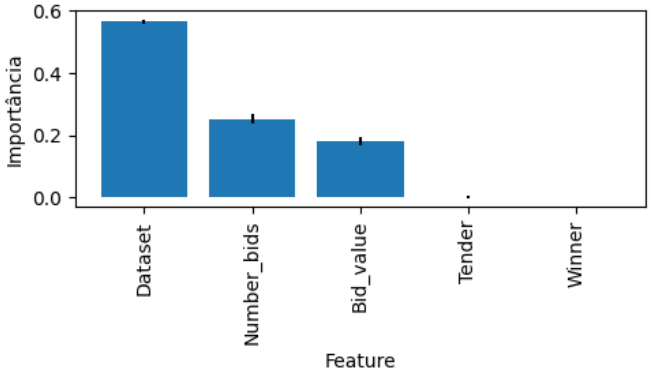


Figure 10. *Feature importance* (conjunto de dados agrupado)



5.4.3. Análise da distribuição dos resultados

Como mencionado anteriormente, o modelo *Logistic Regression* conseguiu diferenciar lances colusivos e não colusivos, mas ajustou-se mais a lances altos, prejudicando o desempenho em lances menores. A Figura 11 mostra múltiplos picos na *feature Number Bids*, com 7 picos corretamente classificados como não colusivos e 6 incorretamente como colusivos. Isso indica que o modelo aprendeu padrões específicos, mas não todos, ao contrário do *Ada Boost* no conjunto de dados brasileiro. Já no conjunto de dados agrupado, as distribuições das *features* numéricas não mostraram diferenças significativas de-

vido às variações nos tipos de licitações e valores. A *feature Number Bids* apresenta uma distribuição bimodal, com maior concentração abaixo de 50 lances, conforme a Figura 12.

Figure 11. Distribuição das *features* (conjunto de dados suíço)

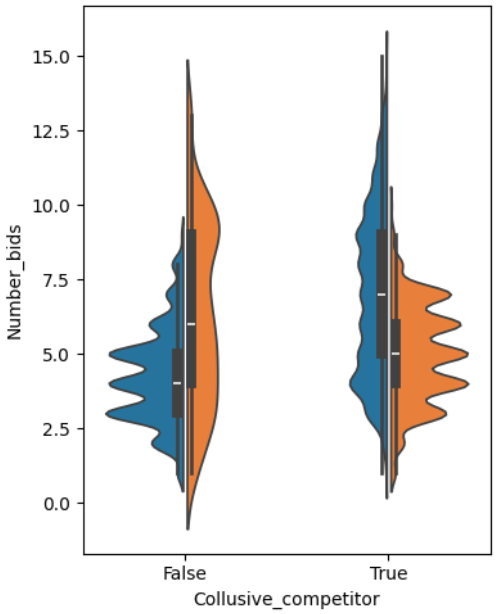
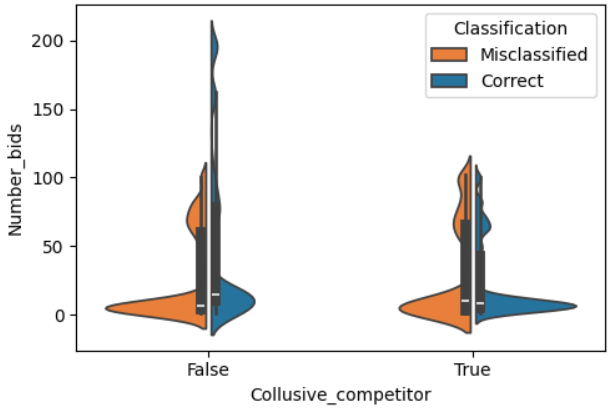
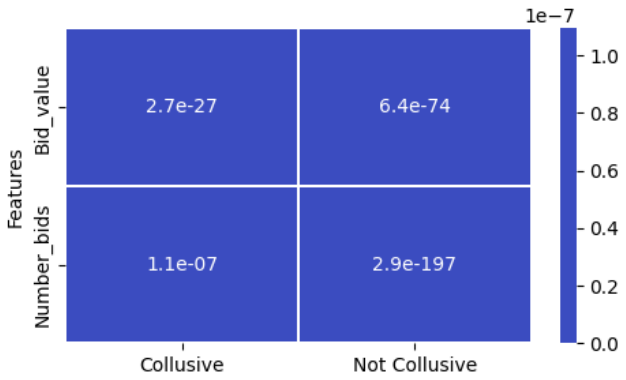


Figure 12. Distribuição das *features* (conjunto de dados agrupado)



Ao analisar os valores de p-value do gráfico de *Mann-Whitney* para o conjunto de dados agrupado na Figura 13, nota-se que há diferença entre as distribuições dos lances classificados corretamente dos incorretos. Ainda que não haja assimetria entre os valores a ponto de facilitar o modelo a encontrar padrões, existe uma variabilidade acentuada, como pode ser observado nas caudas longas das distribuições, fazendo o valor de p-value ser muito baixo e, portanto, haver probabilidade de que as distribuições possuam uma grande diferença.

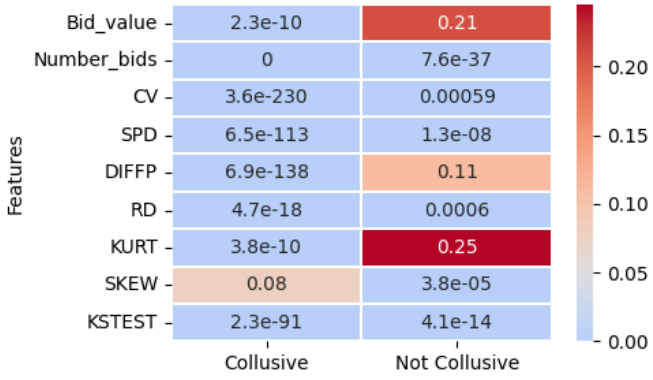
Figure 13. Teste U de *Mann-Whitney* (conjunto de dados agrupado)



No caso do conjunto de dados suíço, evidencia-se melhor na Figura 14 a diferença

entre as distribuições, podendo ser apontada uma similaridade maior nas distribuições de *Bid Value*, *KURT*, *SKEW*, dentre outras. Por outro lado, *features* como *Number bids* possuem valores de p-value igual e tendendo a zero, demonstrando que as distribuições também possuem uma diferença estatisticamente significativa, que se reflete na distribuição da mesma, conforme visto anteriormente.

Figure 14. Teste U de *Mann-Whitney* (conjunto de dados suíço)



6. Conclusão

Neste trabalho, foi estudado acerca da criação de um fluxo de trabalho para auxiliar na tomada de decisão em relação aos melhores modelos de *machine learning* para ser aplicado em conjuntos de dados de conluio em licitações. Para isso, foram realizados experimentos em conjuntos de dados de conluio de diferentes países, sendo eles, Brasil, Suíça e um conjunto de diferentes países agrupados. Com isso, foram reproduzidos os resultados originais com *random subsampling*. Ao observar a variabilidade dos resultados e a falta de informações sobre o desvio padrão reportado pela literatura, foi adaptada a reprodução para serem perfeitamente replicáveis. Em seguida, foram otimizados e selecionados os modelos com maior desempenho, utilizando técnicas de validação cruzada, como *GroupKFold* ou LOGOCV. Por fim, foram analisadas a importância das *features* com base na análise da distribuição das classificações e em testes estatísticos.

Ao realizar a reprodução dos resultados com *random subsampling* para o conjunto de dados brasileiro e posteriormente com LOGOCV, foi observado um aumento no desempenho dos modelos com os hiperparâmetros já disponibilizados pelo artigo de referência, contudo, evidenciou-se um desvio padrão acentuado que, mesmo após realizar uma busca por melhores hiperparâmetros, não foi obtido um resultado capaz de reduzi-lo consideravelmente.

Também constatou-se que este fluxo de trabalho pode auxiliar não somente para a seleção do melhor modelo, como também na análise da qualidade dos dados obtidos. Por meio da análise da variabilidade dos resultados do conjunto de dados brasileiro, constatou-se que a tarefa de encontrar um melhor modelo seria mais desafiadora, sendo necessárias estratégias voltadas a melhorar o conjunto antes de otimizar modelos.

Ademais, o fluxo proposto foi aplicado em outros conjuntos de dados, auxiliando na tomada de decisão em relação à qualidade dos dados e modelos, como demonstrado para conjunto de dados suíço e agrupado. Podemos concluir a partir de ambos que ter uma

maior quantidade de licitações e lances resulta em um desvio padrão menor. Além disso, o uso de modelos mais simples também pode apresentar um desempenho e desvio padrão satisfatórios. Entretanto, é preciso avaliar se as distribuições das classificações feitas pelo modelo estão extraindo padrões em diferentes faixas de valores.

References

Anowar, F. and Sadaoui, S. (2019). Multi-class ensemble learning of imbalanced bidding fraud data. pages 352–358.

OECD (2017). Collusion: Competition policy in the digital age.

Rabuzin, K. and Modrusan, N. (2019). Prediction of public procurement corruption indices using machine learning methods. In *KMIS*, pages 333–340.

Rodríguez, M. J. G., Rodríguez-Montequín, V., Ballesteros-Pérez, P., Love, P. E., and Signor, R. (2022). Collusion detection in public procurement auctions with machine learning algorithms. *Automation in Construction*, 133:104047.

Tan, P.-N., Steinbach, M., and Kumar, V. (2016). *Introduction to data mining*. Pearson Education India.

Velasco, R. B., Carpanese, I., Interian, R., Paulo Neto, O. C., and Ribeiro, C. C. (2021). A decision support system for fraud detection in public procurement. *International Transactions in Operational Research*, 28(1):27–47.

Vera Thorstensen, L. F. G. (2021). Brasil na ocde: Compras públicas.

Wallimann, H., Imhof, D., and Huber, M. (2022). A machine learning approach for flagging incomplete bid-rigging cartels. *Computational Economics*, pages 1–52.

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9):2839–2846.