Previsão da conformidade de projetos de construção pública em Portugal com recurso a aprendizagem automática

https://doi.org/10.21814/uminho.ed.164.36

Luís Jacques de Sousa[1], **João Poças Martins**[2], **Luís Sanhudo**[3]

[1] *CONSTRUCT, FEUP-DEC, 4200-465 Porto; BUILT CoLAB, 4150-003 Porto, 0000-0002-0789-9368*

[2] *CONSTRUCT, FEUP-DEC, 4200-465 Porto; BUILT CoLAB, 4150-003 Porto, 0000-0001-9878-3792*

[3] *BUILT CoLAB, 4150-003 Porto; CONSTRUCT, 4200-465 Porto, 0000-0002-2578-6981*

Resumo

Os projetos de adjudicação públicos são influenciados e decididos de acordo com um conjunto de fatores como o preço base, o prazo de submissão, o número de proponentes, entre outros. Estes fatores podem ter impacto na conformidade orçamental do projeto.

Tradicionalmente a previsão da conformidade orçamental em projetos de construção tem demonstrado ser um grande desafio devido à imprevisibilidade característica de projetos de construção. Não obstante, as técnicas de aprendizagem de máquinas podem oferecer importantes ferramentas de apoio a decisão, através de previsões de conformidade com base em dados históricos.

Aplicações anteriores de aprendizagem de máquinas centraram-se em previsões do custo total da obra com base em dados privados da fase de execução dos projetos, salvo algumas exceções. Neste sentido, este estudo introduz um modelo de aprendizagem de máquinas automática que utiliza dados abertos da fase de adjudicação para prever a conformidade económica de projetos de construção pública. O modelo prevê o cumprimento do orçamento através da análise de diferentes características dos contratos de projetos públicos. Este estudo explora várias arquiteturas de algoritmos e técnicas de tratamento de dados para escolher o modelo com melhor

desempenho com o objetivo de auxiliar o dono de obra na definição os requisitos do concurso.

Ferramentas de aprendizagem de máquinas podem assim fornecer aos donos de obra informações sobre os critérios mais adequados para cada situação com base em projetos semelhantes, ajudando na tomada de decisões. Estudos futuros devem avaliar o impacto e a capacidade do modelo no fluxo de trabalho das adjudicações públicas.

1. Introdução

O processo de adjudicação de obras públicas em Portugal envolve a seriação propostas com base em fatores económicos, temporais e de qualidade. Estas decisões têm impacto significativo no desempenho da obra e são vinculativas contratualmente, ainda assim, estas são muitas vezes feitas em prazos apertados. Neste sentido, ferramentas de apoio à decisão podem auxiliar o dono de obra na escolha do adjudicatário de um projeto de construção [1, 2].

Devido à complexidade e falta de normalização dos documentos de projetos de construção [3], prever com precisão o cumprimento do orçamento nas fases iniciais de projetos tem sido desafiador [4, 5]. No entanto, o desenvolvimento de aplicações com recursos a inteligência artificial pode mudar este paradigma. De facto, estudos recentes no âmbito da indústria AEC têm aplicado diferentes algoritmos de Machine Learning (ML) em diferentes fases do processo de construção [6]. Ainda assim, o setor AEC, quando comparado a setores análogos, é um dos que menos tem adotado estas técnicas [7]. A baixa adoção destas ferramentas deve-se por um lado por razões culturais e técnicas e por outro a vasta diversidade de projetos em Construção [5]. A previsão do orçamento durante a fase de adjudicação é crucial, pois orçamentos mais realistas podem melhorar a eficiência e produtividade, evitando suborçamentação [8].

Estudos anteriores utilizaram diversos modelos de ML para prever orçamentos em projetos de construção [9, 10]. A precisão deste modelos encontra-se geralmente em torno de 90%, ou, com um erro percentual médio absoluto (MAPE) de 20% [11]. A literatura destaca a variabilidade nos dados, indicando que não há correlação evidente entre o número de projetos e o desempenho dos modelos [6]; apesar disso, modelos holísticos exigirão sempre conjuntos de dados extensos devido à natureza dos algoritmos de ML. Para a implementação destes modelos no setor AEC, devem ser recolhidas amostras significativas de dados para treinar os algoritmos [6].

Adicionalmente, a grande maioria das aplicações ML para a previsão de orçamentos de construção utilizaram dados privados da fase de execução da construção [6]. Ainda assim, investigação semelhante recente utilizou informações de contratação acessíveis ao público para prever o preço base de concursos [12] e identificar possíveis colusões em contratos [13].

Analogamente, este estudo desenvolve um modelo de previsão do cumprimento do orçamento em projetos de obras públicas, utilizando dados da fase de adjudicação. O objetivo deste estudo é testar a forma como os algoritmos de previsão de conformidade podem apoiar as decisões dos técnicos de contratação, através da previsão da probabilidade de o projeto cumprir o orçamento.

2. Metodologia

O processo de contratação para um projeto de obras públicas em Portugal inicia após a sua aprovação. Nessa fase, o dono de obra deve determinar características como preço base, prazo de submissão, critério de adjudicação, entre outros fatores, e publicar o anúncio do procedimento em repositórios abertos [14]. Estas características, afetam a forma como as propostas serão elaboradas por parte dos empreiteiros gerais interessados.

Após a seleção do adjudicatário, o projeto de construção é realizado, produzindo resultados financeiros. Estes resultados, bem como a restante informação têm de ser obrigatoriamente submetidos na plataforma de dados aberta Portal Base [15]. Este modelo aproveita a obrigação de reporte em repositórios abertos para usar esses dados como dados de treino.

Como demonstra a Figura 1, este modelo apoia a decisão dos donos de obra relativa a seleção dos critérios de adjudicação do projeto, através da previsão do cumprimento do orçamento, com base no desempenho de projetos anteriores com características semelhantes. Assim, conforme a previsão do modelo, o dono de obra pode alterar ou manter os critérios de adjudicação.

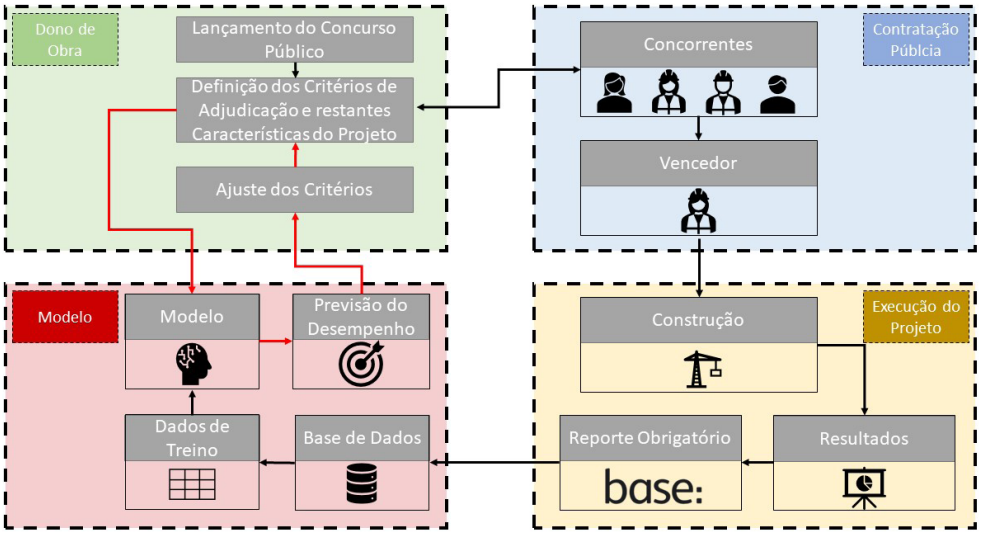

Figura 1
Implementação do modelo no processo de contratação pública.

Todo o código usado para conceber as diferentes arquiteturas testadas foi desenvolvido em Python usando as bibliotecas de ML, Keras e Sklearn [16, 17].

Os seguintes capítulos ilustram a abordagem experimental deste trabalho no qual se insere uma fase de tratamento de dados (incluindo uma análise exploratória de dados e a seleção das caraterísticas), uma fase de testagem de modelos e por fim a discussão dos resultados.

3. Tratamento de dados

3.1. Proveniência e tratamento de dados

A qualidade e quantidade de dados é crucial para o desenvolvimento de aplicações ML e representam o maior desafio conceptual para a implementação destas tecnologias em Construção. No entanto, apenas recentemente a importância da gestão e armazenamento de dados foi reconhecida pelo setor AEC havendo a publicação de alguns trabalhos que partilham bases de dados abertas preparadas para suportar aplicações de ML [18, 19]. Neste sentido, este estudo utilizou dados provenientes da base de dados Portuguese Public Procurement Database (PPPData), que inclui mais de 5000 contratos de públicos provenientes do Portal Base e do Diário da República Eletrónico e que estão caracterizados por 37 propriedades distintas de 2015 a 2022 [18, 19].

Para encontrar um equilíbrio entre a quantidade de dados usados para treinar o algoritmo e o número de caraterísticas optou-se pela seleção das 12 características apresentadas na Tabela 1.

Esta redução do número de características permitiu eliminar apenas 442 contratos devido à exclusão de valores omissos, restando 4772 contratos. A caraterística "Código CPV" exigiu a utilização do codificador LabelEncoder do Sklearn [16] para transformar esta caraterística, de valores não numéricos, em valores numéricos. Além disso, a caraterística "Critério Ambiental" foi transformada numa variável binária (ou seja, 0 e 1) a partir do seu formato lógico anterior. Por último, a caraterística-alvo “Desempenho” foi adaptada a um formato categórico. Não foram necessários esforços adicionais de normalização, uma vez que a base de dados já se encontrava num formato estruturado.

Tabela 1: Caraterísticas do modelo

Nome	Descrição	Formato	Unidade de Medição
Código CPV	Código "Common procurement vocabulary". Exclusivamente códigos de Construção	String	N/A
Critério Ambiental	Se o critério ambiental foi considerado no concurso (TRUE-FALSE)	Lógico	N/A
Ano de Publicação	Ano de publicação do concurso no Portal Base	Data	N/A
Distrito	Local de execução do projeto. Código de identificação do distrito, organizado alfabeticamente e numerado de 1 a 20	Inteiro	N/A
Prazo de Submissão	Prazo para a aceitação dos pedidos de propostas	Inteiro	Dias
Prazo de Execução	Prazo para a conclusão do projeto	Inteiro	Dias
Preço Base	O montante máximo que o cliente está disposto a pagar pela execução do projeto.	Moeda	€
Preço Inicial	O preço inicial acordado entre o dono de obra e o vencedor do concurso.	Moeda	€
Categoria do Preço Inicial	Classe atribuída ao preço inicial: (1) – Entre 0 e 250 mil, (2) – Entre 250 mil e 1 milhão, (3) – Mais de 1 milhão	Inteiro	N/A
Critério de Adjudicação	Os critérios de adjudicação utilizados para selecionar o vencedor, tal como indicado nos documentos do concurso. (1) – Critério Multifactor, (2) – Critério do preço mais baixo, (0) – Critério de adjudicação em falta	Inteiro	N/A
Peso do Preço no Critério de Adjudicação	A ponderação dada ao fator preço nos critérios de adjudicação	Percentagem	N/A
Número de Concorrentes	Número de proponentes no concurso	Inteiro	N/A
Nome da caraterística alvo	**Descrição**	**Formato**	**Unidade de Medição**
Desempenho	A classe atribuída ao resultado do projeto: Deslize de preço: Preço Efetivo/ Inicial >= 105% Conformidade de preço: 95% < Preço Efetivo/ Inicial < 105% Poupança de preço: Preço Efetivo/ Inicial =< 95%	Inteiro	N/A

Adicionalmente, estudos estatísticos aplicados ao PPPData demonstram um grande desequilíbrio relativamente à classe "Desempenho". De facto, depois do tratamento dos dados, verificou-se essa mesma distribuição desequilibrada, com a classe de conformidade de preços a registar 3290 contratos, a de deslize de preços a registar 794 contratos e a de poupança de preços a registar 688 contratos. Assim, para equilibrar este conjunto de dados e atenuar a influência da assimetria nos resultados, foram testadas três técnicas de balanceamento diferentes: Oversampling [20]; Class weights balancing; Synthetic Minority Oversampling Technique (SMOTE) [21]. Assim, foi efetuado um teste utilizando o algoritmo Adam [22] para comparar as diferentes técnicas de balanceamento de dados. Foi definida uma configuração fixa para o algoritmo, a fim de os comparar a sua precisão e eficiência entre cada um dos conjuntos de dados equilibrados. Embora não exista uma configuração universalmente

estabelecida para este modelo, umas das configurações mais utilizadas neste algoritmo e a usada para este teste incluiu uma divisão treino-teste de 0,33, um tamanho de série ("batch") de 32 e 50 épocas ("epochs"). A Tabela 2 apresenta os resultados.

Tabela 2: Métricas de performance dos testes de balanceamento de dados

Técnica	Médias Macro			Loss de Validação	Exatidão de Validação	Tempo de Computação (s)
	Precisão	Recall	F1-score			
Oversampling	0.75	0.77	0.76	0.29	0.82	9.9
Class Weights	0.82	0.70	0.74	0.36	0.76	11.2
SMOTE	0.85	0.85	0.85	0.25	0.85	19.6

A partir da Tabela 2 percebe-se que a estratégia de Class weeights balancing é a técnica com os piores resultados. As técnicas de Oversampling e SMOTE obtiveram resultados semelhantes; no entanto, apesar de o método de Oversampling efetuar as 50 epochs em menos tempo do que o método SMOTE, este teve uma melhoria de 0,03 na precisão da validação. Relativamente aos resultados das médias macro, o método SMOTE obteve os melhores resultados entre as três técnicas.

Com o intuito de esclarecer qual dos dois métodos se adapta melhor para resolver este problema, foi analisada a precisão ao longo do tempo, conforme apresentado na Figura 2.

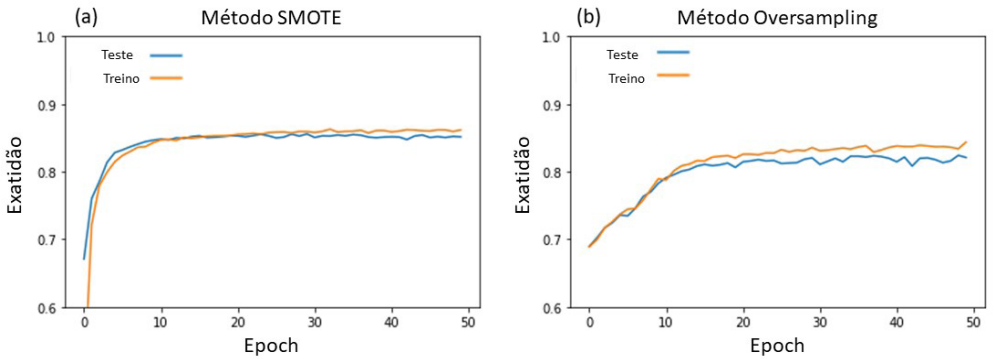

Figura 2
Desempenho do método SMOTE vs Oversampling.

Comparando a curva da Figura 2a e 2b, verifica-se que o método SMOTE obteve uma curva de treino suave, indicando maior estabilidade comparando com método Oversampling. Além disso, o SMOTE atinge uma exatidão máxima mais cedo do que o método de Oversampling (aproximadamente 85% de exatidão de validação em dez epochs, em comparação com 82% de exatidão de validação em 22 epochs para o método de Oversampling). Devido a todos estes fatores, o método SMOTE foi aplicado para equilibrar o conjunto de dados deste problema.

3.2. Análise de dados exploratória e seleção de características

Após o tratamento dos dados, de forma a perceber quais variáveis se sobrepõem umas às outras e quais são irrelevantes para a previsão e podem ser descartadas efetuaram-se dois testes: um teste de correlação e uma análise de importância de caraterísticas.

A Figura 3 apresenta um mapa de correlação entre as caraterísticas. Quando a correlação é próxima de 1 ou -1, é alta; quando é próxima de 0, é baixa. Os números negativos representam correlações inversas. O mapa mostra que existem grupos de caraterísticas com correlações significativas no conjunto de dados equilibrado. Por exemplo, existe uma correlação elevada entre o "Critérios de Adjudicação" e o "Peso do Preço no Critério de Adjudicação" (0,99). Além disso, as características "Preço base" e "Preço inicial" também estão intimamente relacionadas (0,98). A Figura 3 mostra ainda que o "Desempenho" está principalmente relacionado com o "Ano de Publicação", o "Critério de Adjudicação" e o "Peso do Preço no Critério de adjudicação". Além disso, as características menos correlacionadas com o "Desempenho" são o "Distrito" e o "Número de Concorrentes".

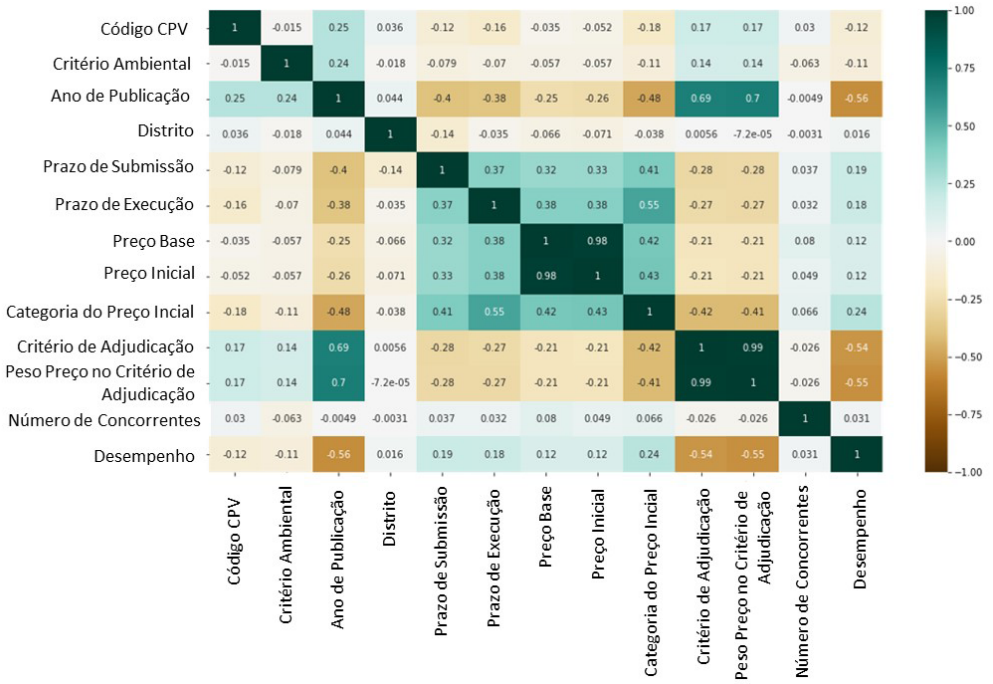

Figura 3
Mapa de correções entre variáveis.

A análise exploratória dos dados permitiu a seleção das características a considerar para o modelo. No entanto, esta análise não deve ser considerada de forma exclusiva, assim, para compreender melhor o impacto das caraterísticas no modelo, foi efetuada uma análise da importância de caraterísticas utilizando o método Random Forest [23], nomeadamente, a importância com base na impureza e a importância através da permutação [24].

Neste estudo, o algoritmo Random Forest do Scikit-learn foi implementado para obter a importância das caraterísticas utilizando a classe RandomForestClassifier. Depois de treinar o modelo, a classe fornece uma propriedade incorporada (feature_ importances_) que define a pontuação de importância relativa de cada caraterística de entrada.

Analogamente, a importância da caraterística através da permutação utilizou o classificador Random Forest para obter a importância da caraterística através da permutação dos dados de treino com o conjunto de validação. A análise das características através da permutação permite calcular as pontuações de importância relativa independentemente do modelo utilizado, o que é vantajoso do ponto de vista da comparação com o método baseado na impureza.

Os resultados da análise da importância das características são apresentados na Figura 4, onde as características mais relevantes são ordenadas da esquerda para a direita e uma linha de limite (a vermelho), fixada a 95%, foi colocada de modo a intersectar a soma cumulativa da importância das características mais relevantes. Ambos os métodos destacam o "Ano de Publicação" como a caraterística de maior impacto para a previsão. O "Critérios de atribuição" foi a segunda caraterística mais relevante para a classificação em ambos os testes de importância. A ordem de importância das restantes características difere de um método para outro; apesar disso, algumas características podem ser selecionadas por serem relevantes em ambos os testes. Entre estas características, o "Peso do Preço no Critério de Adjudicação", o "Preço inicial", o "Preço Base", o "Número de Concorrentes", o "Prazo de Submissão" e "Prazo de Execução" são destacados como significativos em ambos os testes.

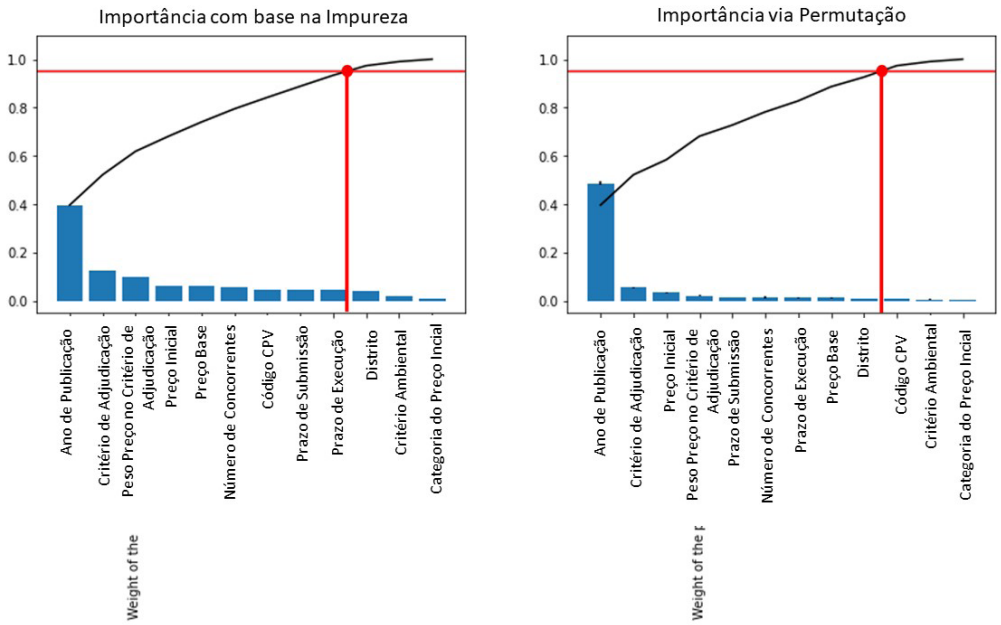

Figura 4
Resultados da análise de importância das variáveis.

Tendo em conta as correlações entre variáveis, a análise de importância das características e a natureza do processo de contratação da construção, foram selecionadas oito características entre as doze disponíveis. O "Ano de Publicação" foi selecionado por ter sido considerado o elemento com maior impacto. O “Critério de Adjudicação" e o "Preço Base" foram escolhidos por poderem ser personalizados no momento da publicação do anúncio de concurso, em detrimento do "Peso do Preço no Critérios de Adjudicação" e do "Preço Inicial", por estarem significativamente correlacionados. Por fim, o "Prazo de Submissão", o "Prazo de Execução", o "Número de Concorrentes", o "Código CPV" e o "Distrito" completaram o grupo de características selecionadas, descartando as restantes.

4. Seleção de modelos

Uma revisão da literatura sobre a utilização do ML para prever o cumprimento do orçamento não mostra consenso sobre qual tipo de algoritmo conduz a resultados ótimos [6]. Os resultados de cada arquitetura estão muito relacionados com os dados que compõem o modelo. Para tal, devem ser testados diferentes algoritmos de modo a encontrar a melhor solução para cada tipo de aplicação. Neste contexto, foram testados os seguintes algoritmos de forma a selecionar o que melhor se adequa a este estudo: (1) Adam ANN; (2) Random Forest; (3) SVM; (4) Extreme Gradient Boosting (Xgboost);(5) K-nearest Neighbours (KNN).

Todos os algoritmos foram desenvolvidos em Python, utilizando as bibliotecas Keras e Sckit-learn. Todos os métodos foram submetidos a rigorosos estudos de experimentação para obter a melhor configuração possível. A Tabela 3 demonstra as definições de cada algoritmo na sua configuração com melhor desempenho e compara os resultados de cada modelo. Todos os modelos utilizaram um train/test split de 0,33.

Tabela 3: Resultados do Teste dos Algoritmos

Algoritmo	**Hiperparâmetros**	**Resultados**				
		Exatidão de Treino	**Exatidão de Teste**	**Precisão**	**Recall**	**F1--Score**
Adam ANN	Uma input layer, uma hidden layer e uma output layer, Input layer with 8 nodes, Hidden layer with 16 nodes, Output layer with 3 nodes, Kernel initialiser = Glorot uniform, Activation function = Rectified Linear Unit (RELU), Batch size = 32, Epoch = 100. (Plateau depois de 40)	0.875	0.855	0.853	0.852	0.851
Random Forest	GridSearchCV: Bootstrap = False, Maximum depth = 4, Maximum features = sqrt, Minimum samples leaf' = 2, Minimum samples split = 2, Number of estimators = 33.	0.842	0.832	0.830	0.866	0.846
SVM	RandomSearchCV, C=10, Gamma=1, Kernel = Radial Basis Function	0.982	0.884	0.883	0.882	0.882
Xgboost	GridSearchCV: Booster = gbtree, Maximum delta step = 0, Maximum depth = 9, Minimum child weight = 1, Number of estimators = 64, Sampling method = uniform	0.999 (sobrea-juste)	0.896	0.917	0.909	0.913
KNN	GridSearchCV: metric='euclidean' n_neighbors=1000, weights='distance'	1 (sobrea-juste)	0.750	0.748	0.766	0.719

Como visto na Tabela 3, o método KNN teve um desempenho fraco, apresentando sobreajuste aos dados durante o treino e resultados de validação baixos, o que levou à sua exclusão. Do mesmo modo, os métodos XGBoost e SVM revelaram tendência para o sobreajuste e foram igualmente excluídos. Os modelos Adam e Random Forest obtiveram uma exatidão de validação de cerca de 85%, com o algoritmo Adam a ter um desempenho ligeiramente superior.

Numa análise secundaria foi excluída a caraterística "Ano de Publicação" e a precisão dos dois melhores algoritmos da análise anterior foi testada. Esta medida foi tomada por duas razões: (1) para avaliar a influência desta variável no desempenho do algoritmo e (2) porque o dono de obra não pode personalizá-la durante a configuração dos critérios do concurso público.

Como visto na Tabela 4, a remoção desta variável resultou numa queda de 20% na precisão de ambos os algoritmos. Apesar disso, o Adam continuou a ter o melhor desempenho e o Random Forest mostrou forte consistência entre as fases de treino e validação. Ambos os algoritmos atingiram uma precisão de aproximadamente 65%, que foi considerada realista para algoritmos de apoio à configuração de concursos públicos.

Tabela 4: Resultados do 2.º Teste de Algortimos

Algoritmo	**Hiperparâmetros**	**Resultados**				
		Exatidão de treino	**Exatidão de Teste**	**Precisão**	**Recall**	**F1--Score**
Adam ANN	Uma input layer, uma hidden layer e uma output layer, Input layer with 7 nodes, Hidden layer with 14 nodes, Output layer with 3 nodes, Kernel initialiser = Glorot uniform, Activation function = Rectified Linear Unit (RELU), Batch size = 32, Epoch = 100. (Plateau depois de 40)	0.679	0.683	0.681	0.675	0.671
Random Forest	GridSearchCV: Bootstrap =True, Maximum depth = 4, Maximum features = sqrt, Minimum samples leaf' = 1, Minimum samples split = 2, Number of estimators = 10.	0.622	0.618	0.615	0.764	0.660

5. Discussão dos resultados e conclusão

Este estudo visou criar um modelo de previsão de conformidade financeira para contratos públicos em Construção, empregando o PPPData, aplicando o caso de estudo a contratos em Portugal. O método SMOTE foi escolhido para equilibrar o conjunto de dados.

Uma análise estatística das variáveis de entrada não revelou correlação direta entre a escolha de um critério de adjudicação específico e o cumprimento do orçamento. Além disso, não houve correlação identificada entre a extensão dos prazos de apresentação e a melhoria do desempenho do projeto. Essas conclusões destacam a complexidade da previsão do orçamento nas fases iniciais do projeto, dada a

variabilidade no desempenho dos contratos de construção, o que constitui um desafio significativo.

Doze variáveis, identificadas através de análises de correlação e de importância, orientam o treino do modelo. Os algoritmos Adam ANN e Random Forest foram identificados como adequados ao problema de classificação com uma precisão de validação de cerca de 85%.

No segundo teste, excluindo a variável "Ano de Publicação", o algoritmo Adam ANN demonstrou o melhor desempenho para este problema, estabelecendo uma precisão de referência de 68% para ferramentas de apoio à decisão em concursos públicos na construção.

Os resultados deste estudo provam que o ML não consegue prever com precisão a taxa de conformidade dos projetos de construção. Embora os resultados dos algoritmos possam nem sempre ser mais precisos do que os obtidos por humanos, a vantagem notável reside na sua capacidade de fornecer previsões numa fração de tempo em comparação com um técnico. Além disso, proporciona aos decisores justificações com base em dados históricos. Este aspeto sublinha a sua eficácia como ferramentas valiosas durante o processo de definição dos critérios de aquisição

Os estudos futuros devem procurar expandir a base de dados com dados provenientes de outros países, aumentar o número de características de entrada, melhorar a precisão e analisar os impactos deste modelo na eficiência destes processos.

Referências

[1] M. S. A. Aman and S. Azeanita, "Building Information Modelling for Project Cost Estimation," *Recent Trends in Civil Engineering and Built Environment,* vol. 3, no. 1, pp. 621-630, 12/04 2021.

[2] S. Moon, S. Chi, and S.-B. Im, "Automated detection of contractual risk clauses from construction specifications using bidirectional encoder representations from transformers (BERT)," *Automation in Construction,* vol. 142, p. 104465, 2022/10/01/ 2022, doi: https://doi.org/10.1016/j.autcon.2022.104465.

[3] J. P. d. S. P. Martins, "Modelação do fluxo de informação no processo de construção: aplicação ao licenciamento automático de projectos", no. Porto, 2009.

[4] P. Jafari, M. Al Hattab, E. Mohamed, and S. Abourizk, "Automated extraction and time-cost prediction of contractual reporting requirements in construction using natural language processing and simulation," *Applied Sciences (Switzerland),* Article vol. 11, no. 13, 2021, Art no. 6188, doi: 10.3390/app11136188.

[5] H. Elhegazy *et al.*, "Artificial Intelligence for Developing Accurate Preliminary Cost Estimates for Composite Flooring Systems of Multi-Storey Buildings," *Journal of Asian Architecture and Building Engineering,* 2021, doi: 10.1080/13467581.2020.1838288.

[6] L. Jacques de Sousa, J. Poças Martins, J. Santos Baptista, and L. Sanhudo, "Towards the Development of a Budget Categorisation Machine Learning Tool: A Review," in *Trends on Construction in the Digital Era*, Guimarães, Portugal, A. Gomes Correia, M. Azenha, P. J. S. Cruz, P. Novais, and P. Pereira, Eds., 2023// 2023: Springer International Publishing, pp. 101-110, doi: https://doi.org/10.1007/978-3-031-20241-4.

[7] D. Chen, L. Hajderanj, and J. Fiske, "Towards automated cost analysis, benchmarking and estimating in construction: A machine learning approach," in *Multi Conference on Computer Science and Information Systems, MCCSIS 2019 – Proceedings of the International Conferences on Big Data Analytics, Data Mining and Computational Intelligence 2019 and Theory and Practice in Modern Computing 2019*, 2019, pp. 85-91, doi: 10.33965/bigdaci2019_201907l011. [Online].

[8] S. Kerridge and C. Halaris, *SupplyPoint: An integrated system supporting E-business in the Construction Sector*. 2001.

[9] N. Suneja, J. P. Shah, Z. H. Shah, and M. S. Holia, "A neural network approach to design reality oriented cost estimate model for infrastructure projects," *Reliability: Theory and Applications,* Article vol. 16, pp. 254-263, 2021. [Online].

[10] X. Y. Jiang, N. Y. Pa, W. C. Wang, T. T. Yang, and W. T. Pan, "Site Selection and Layout of Earthquake Rescue Center Based on K-Means Clustering and Fruit Fly Optimization Algorithm," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 27-29 June 2020 2020, pp. 1381--1389, doi: 10.1109/ICAICA50127.2020.9182505.

[11] M. Juszczyk, "Development of Cost Estimation Models Based on ANN Ensembles and the SVM Method," *Civil And Environmental Engineering Reports,* vol. 30, no. 3, pp. 48-67, 2020, doi: 10.2478/ceer-2020-0033.

[12] M. J. García Rodríguez, V. Montequín, F. Ortega-Fernández, and J. Balsera, "Bidders Recommender for Public Procurement Auctions Using Machine Learning: Data Analysis, Algorithm, and Case Study with Tenders from Spain," *Complexity,* vol. 2020, 11/25 2020, doi: 10.1155/2020/8858258.

[13] M. J. García Rodríguez, V. Rodríguez-Montequín, P. Ballesteros-Pérez, P. E. D. Love, and R. Signor, "Collusion detection in public procurement auctions with machine learning algorithms," *Automation in Construction,* vol. 133, p. 104047, 2022/01/01/ 2022, doi: https://doi.org/10.1016/j.autcon.2021.104047.

[14] R. Portuguesa. "Diário da República Electrónico." https://dre.pt/dre/home (accessed 2023).

[15] I. d. M. P. d. I. e. d. C. (IMPIC). "Portal Base." https://www.base.gov.pt/ (accessed 2023).

[16] "Scikit-Learn." https://scikit-learn.org/ (accessed August 2023.

[17] "Keras." https://keras.io/ (accessed August 2023.

[18] L. Jacques de Sousa, J. Poças Martins, and L. Sanhudo, "Base de dados: Contratação pública em Portugal entre 2015 e 2022," presented at the Construção 2022, Guimarães, Portugal, 2022.

[19] L. Jacques de Sousa, J. Poças Martins, and L. Sanhudo, "Portuguese public procurement data for construction (2015-2022)," *Data in Brief,* vol. 48, p. 109063, 2023/06/01/ 2023, doi: https://doi.org/10.1016/j.dib.2023.109063.

[20] J. Brownlee. "Tour of Data Sampling Methods for Imbalanced Classification." https://machinelearningmastery.com/data-sampling-methods-for-imbalanced-classification/#:~:text=The%20simplest%20oversampling%20method%20involves,for%20Synthetic%20Minority%20Oversampling%20Technique. (accessed 2023).

[21] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics,* vol. 14, no. 1, p. 106, 2013/03/22 2013, doi: 10.1186/1471-2105-14-106.