

EU Public Procurement Contract Value Prediction: A Machine Learning Approach

Dr Muhammad Azizur Rahman, Anna Hlinska

Department of Computer Science, Cardiff Metropolitan University, Llandaff Campus, Western Avenue, Cardiff, CF5 2YB, UK

Abstract—Using modern analytical tools, such as machine learning, organisations can collect and analyse large amounts of data related to suppliers, pricing, demand structure, and market trends. This study uses machine learning models to predict the value of EU public procurement contracts based on the Tenders Electronic Daily (TED) database. The analysis covers 13,345,120 initial contract records for the period 2006-2025 in 33 states and 63 procurement sectors. After careful data quality control procedures, the analytical set was formed by 10,038,018 valid contracts (75% retention rate). Three complementary methodologies were used: Random Forest regression to identify nonlinear patterns, ordinary least squares (OLS) regression to interpret coefficients, and K-means clustering to classify procurement behaviour at the country level. The Random Forest achieved cross-validation $R^2=0.2795$ and the test $R^2=0.2613$, with the country of origin dominating the predictive value (Germany: 24.39%, United Kingdom: 2.33%, Italy: 1.96%). Temporal features had a significance of 18.36%, while competition indicators (number of proposals: 8.99%) and structural characteristics (batch number: 12.29%) had a significant impact. The OLS regression showed statistically significant effects for countries: Germany showed 98.4% lower contract prices despite being Europe's largest economy, reflecting federal administrative fragmentation, while Italy showed 295.7% higher rates thanks to centralised infrastructure projects. The K-means clustering revealed three clear procurement profiles: Greece as a transparency-focused outsider (109 average bids per contract), 19 mature economies with high value, and 13 less valuable, fragmented systems, including Germany. The results show that institutional frameworks dominate economic factors in determining the value of contracts, which has political implications for the design of procurement systems across the EU.

Keywords—Public procurement; EU procurement; machine learning; Random Forest; Contract Value Prediction; TED database; institutional economics; OLS regression; K-means clustering; predictive modelling; data quality; EU AI Act

I. INTRODUCTION

Public procurement constitutes a significant component of public expenditure in the member states of the European Union, accounting for approximately 14% of GDP or 2 trillion euros annually [1]. The Tenders Electronic Daily (TED) database serves as the central repository for procurement information, publishing contract award notices that provide transparency into public spending decisions across 33 EU and European Economic Area countries. Despite this wealth of data, predicting contract values remains challenging due to complex interactions between country-specific institutional frameworks, sectoral characteristics, temporal trends, and competitive dynamics [2, 3]. Applying machine learning to public procurement data,

budget planning, supplier strategies, and procurement policy evaluation. The scale and complexity of EU public procurement present both opportunities and challenges for data-driven analysis. Accurate contract value forecasting has significant policy and practical implications for many stakeholder groups. Procurement agencies need reliable forecasts for budget planning and resource allocation, especially against the backdrop of increased fiscal constraints in European governments. Suppliers benefit from understanding typical contract sizes when evaluating market entry decisions in different countries and sectors, which has implications for competitive strategy and capacity planning [5]. Traditional approaches to contract value analysis primarily rely on descriptive statistics and linear regression models, which may not consider the nonlinear relationships and interaction effects characteristic of procurement processes [6].

Recent advances in machine learning provide opportunities to improve prediction accuracy and uncover hidden patterns through algorithms capable of detecting complex relationships in high-dimensional datasets [7, 8]. The practical challenges of applying AI to procurement data go beyond algorithmic selection and encompass the fundamental limitations of data infrastructure. Recent comprehensive assessments show that while all EU Member States now publish structured procurement data that meets minimum machine learning standards, critical quality issues remain: approximately 30% of TED recordings contain missing values – this figure has not changed since 2017 – with supplier registration numbers missing in 86% of procedures and evaluation contracts missing in 63% of cases [22].

This study examines several critical gaps in the existing literature. First, it provides comprehensive machine learning analysis of the value of EU public procurement contracts, spanning nearly two decades (2006 – 2025) and covering all Member States, going beyond the country-specific or sectoral analyses that dominate modern research. Second, it systematically compares multiple modelling approaches – Random Forest for prediction accuracy, OLS regression for interpretive coefficients, and K-means clustering for identifying behavioural patterns — to provide both predictive capability and essential insights into the factors that determine contract value. Third, it quantifies the relative importance of country-level institutional factors compared to sectoral and temporal characteristics, which provides empirical evidence for the ongoing debate on the role of institutional quality in public procurement outcomes [12]. Fourth, it demonstrates the practical application of machine learning techniques to real-

world procurement data, which is characterised by missing values, inconsistent formats, and excessive computational loads.

The rest of this study is organised as follows: Section II reviews related work. Section III describes the research methodology. Section IV presents the results and analysis. Section V discusses the findings. Section VI outlines future directions, and Section VII concludes the study.

II. RELATED WORK

The literature review summarises the key findings of previous studies, organised by geographical coverage, methodological approach, and content focus. Table I provides a comparative overview of key prior studies alongside this work. The latest comprehensive assessments of procurement data infrastructure show significant progress along with ongoing challenges. Longobucco and Ferwerda (2025) [22] conducted a systematic assessment of the suitability of procurement data for AI-based anti-corruption analysis in all EU Member States and associated countries. Their "funnel logic" system assesses data availability at several levels: structured procurement data available en masse, beneficial ownership registers, political finance disclosures, Politically Exposed Persons (PEP) databases, procurement complaints, and media ownership records. The study confirms that all EU-27 countries are now publishing procurement data that meets minimum standards for machine learning applications, with Croatia, Estonia, and Latvia recognised as leaders in data quality and availability. However, critical gaps remain: approximately 30% of TED recordings contain missing values — this figure has not changed since 2017 — with national vendor registration numbers missing in 86% of procedures and missing contract evaluation numbers in 63% of cases [22]. These systemic data quality issues directly limit the predictive performance achieved by even complex algorithms, setting realistic expectations for model accuracy [10]. García Rodríguez et al. (2021) [11] analysed 58,337 Spanish tenders between 2012 and 2018, comparing linear regression, Random Forest, isotonic regression, and artificial neural networks to predict reward prices. Their conclusion that neural networks and isotonic regression outperformed Random Forest differs from

our results, likely reflecting their focus on a single country's relatively homogeneous institutional environment, dominated by nonlinear sector-specific patterns, in contrast to the pan-European sample, where institutional heterogeneity in 33 countries contributes to Random Forest's ability to capture country-specific effects [17, 18]. Recent surveys of AI and ML applications in procurement [19], combined with country-specific studies in Thailand [20] and Portugal [21], demonstrate comparable levels of accuracy, but their narrower scope limits generalisation to the pan-European institutional diversity examined here.

This research goes beyond existing research in several dimensions. First, it provides the most comprehensive geographical and temporal coverage to date: 10 million valid contracts for the period 2006–2025 in all 33 EEA and EEA member states, compared to previous studies targeting individual countries or sectors. Second, it systematically addresses the data quality issues documented by Longobucco and Verwerda (2025) [22] through explicit preprocessing procedures: deletion of 21.6% of records with missing contract values, exclusion of exceptions over €100 million, standardisation of incompatible formats, and implementation of stratified sample selection to maintain representativeness despite computational limitations. Third, it combines predictive modelling (Random Forest), causal inference (OLS regression with interpreted coefficients), and unsupervised pattern detection (K-means clustering), allowing triangulation between multiple analytical frameworks. Fourth, it quantifies the priority of institutional over economic determinants through the analysis of the importance of features, directly testing competing theoretical views on procurement management. Finally, it establishes a methodological framework aligned with the new standards for AI governance in the EU (EU Regulation 2024/1689 on the AI Law 2024/1689, MCC-AI guidelines), demonstrating how government organisations can develop sophisticated analytical capabilities using open algorithms and public datasets while maintaining transparency and vendor independence [15, 16].

TABLE I. COMPARISON OF CONTRACT VALUE PREDICTION STUDIES

Ref	Geographic Scope & Data	Approach	Algorithm(s)	Performance
[11]	Spain, 58,337 (2012–2018)	Supervised	LR, RF, Isotonic, ANN	ANN & Isotonic > RF
[19]	Multiple countries, 2024 review	Supervised	Mixed methods review	Systematic review
[20]	Thailand, 283,000 (2019)	Supervised	ANN, DT, KNN	77.8% accuracy
[21]	Portugal construction, 2015–2022	Supervised	SVM, RF, +3 others	Multiple compared
This	EU-wide 33 states, 10M (2006–2025)	Supervised + Unsupervised	RF, OLS, K-means	RF R ² =0.2613 / OLS R ² =0.1552

III. RESEARCH METHODOLOGY

The main objective of this study is to predict the value of contracts in EU public procurement using machine learning methods and to identify the institutional determinants that most significantly affect these values. The methodology is structured in four main stages: data collection and pre-processing, feature engineering, model development, and evaluation.

A total of 13,345,120 procurement contracts were downloaded and extracted from the Tenders Electronic Daily

(TED) database, covering the period from 2006 to 2025 in all 33 EEC and EEA member states. TED is the official source of European public procurement reports, providing comprehensive data on contracts exceeding EU thresholds [14]. After rigorous data cleaning and validation — including the removal of records with missing contract values (21.6% of the original dataset), the elimination of exceptions exceeding €100 million, and the exclusion of incomplete records — the final dataset contained 10,038,018 valid contracts, representing a retention rate of 75%. This significant dataset provides reliable coverage of 63

procurement sectors classified under the Common Procurement Vocabulary (CPV) system, ensuring representativeness in both geographical and sectoral dimensions. These values of lack of value are closely aligned with the pan-European data quality assessments, which document approximately 30% of incomplete TED recordings, with 21-30% of cases missing contract values, and 86-89% of procedures lacking buyer information [22], confirming that problems with observed data reflect systemic infrastructure limitations rather than extraction errors.

The data extraction process used the TED API and bulk upload capabilities. The key variables that stand out are: contract value (dependent variable), country of the contracting authority, year of contract, procurement sector (CPV code), batch number within framework agreements, number of proposals received, and details of the buyer's organisation. Due to computational constraints and the need for balanced model training, a stratified random sample of 10,000 observations was taken from a cleaned dataset, which maintained proportional representation across countries, sectors, and time periods. Fig. 1 presents the methodology framework.



Fig. 1. Methodology framework.

IV. RESULTS AND ANALYSIS

This section presents the empirical findings of our machine learning analysis on 10,000 EU public procurement contracts. We look at four key dimensions: (A) the characteristics and preparation of the dataset, (B) the importance of features and performance of the comparative visualisation model, (C) the impacts for individual countries and their institutional drivers, and (D) clustering patterns in European procurement systems.

A. Dataset Characteristics

The results are presented through tables (Table II to Table VI) and figures (Fig. 2 to Fig. 6), which together demonstrate the dominance of institutional factors over economic determinants. Table II summarises the key characteristics of the final analytical dataset.

TABLE II. CHARACTERISTICS OF THE DATASET

Specifications	Cost
Total number of contracts	10,000
Geographical coverage	33 European countries
Time slot	2005-2023 (19 years old)
Training kit	8,000 (80%)
Test Kit	2,000 (20%)
Total number of traits (after coding)	128
Features of the country	33
Features of the sector	18
Average contract value	€9,173,061
Median contract value	€545,420

Fig. 2 shows the temporal evolution, showing steady growth from 2005 to the peak of 2021 (1,092 contracts) and then normalising to 966 in 2023. The sectoral distribution (Fig. 3) shows the dominance of medical equipment at the level of 37% (3,695 contracts). Poland dominates the sample with 22% (2,198 contracts), while Germany has only 7% (705 contracts), despite its economic size.

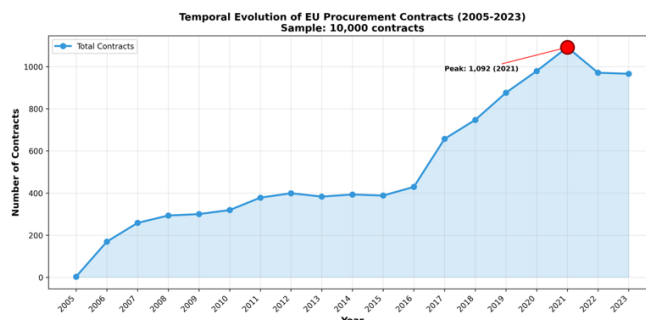


Fig. 2. Temporary evolution of EU procurement contracts (2005-2023).

Medical Equipment dominates at 37.0%, while Unknown (31.0%) reflects incomplete CPV coding in the TED database, especially for pre-2010 contracts.

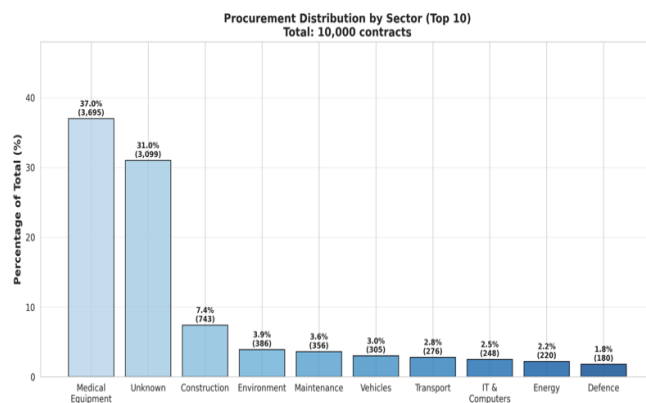


Fig. 3. Distribution of purchases by sectors.

Poland dominates the sample with 22% (2,198 contracts), followed by France (15.7%) and Romania (13.6%). Germany, despite its economic size, represents only 7% of the sample, which implies the fragmentation effects identified in section D.

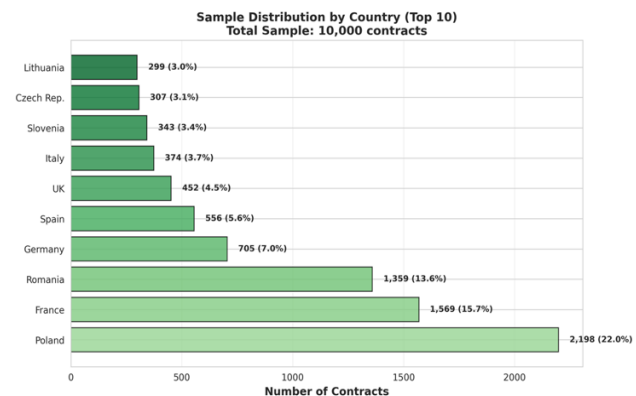


Fig. 4. Distribution of the sample by country (top 10).

B. Model Performance Comparison

This section presents the empirical findings of three complementary modelling approaches: Random Forest regression for [9] forecasting accuracy, ordinary least squares (OLS) regression to estimate interpreted coefficients, and K-means clustering to classify country-level procurement behaviour. The results are organised by model performance, feature importance analysis, country influences, and clustering patterns.

TABLE III. MODEL PERFORMANCE COMPARISON

Model	R ² (CV)	R ² (Test)	RMSE	Std Dev
Random Forest	0.2795	0.2613	2.549	±0.0212
OLS regression	—	0.1552	2.726	N/A
K-average clustering	—	3 clusters	—	N/A

Random Forest outperforms OLS by 68.4% in R² (0.2613 vs. 0.1552), demonstrating better predictive ability. A deviation of the cross-validation standard ±0.0212 indicates stable performance in warehouses with minimal overrun (flow R²=0.4366, test R²=0.2613, gap=0.1753).

C. Feature Importance Analysis

The feature importance analysis was conducted on the test set to identify the most influential predictors. Table IV presents the top 15 features ranked by their contribution to model variance [4].

The analysis reveals three key findings. First, Germany (country_DE) alone represents 24.39% of the predictive power, despite representing only 7% of the sample, suggesting an extraordinary institutional impact on surveillance.

TABLE IV. FEATURE IMPORTANCE RANKINGS (TOP 15)

Rank	Feature	Importance (%)	Category
1	Country: Germany (DE)	24.39	Institutional
2	Year	18.36	Temporal
3	Lot Number	12.29	Structural
4	Number of Offers	8.99	Competition
5	Month	6.14	Temporal
6	Sector: Medical Equipment	2.78	Economic
7	CPV Group 33	2.49	Classification
8	Country: United Kingdom (UK)	2.33	Institutional
9	Country: Italy (IT)	1.96	Institutional
10	Country: Netherlands (NL)	1.71	Institutional
11	Procedure Type: W	1.70	Procedural
12	Country: France (FR)	1.19	Institutional
13	Sector: Unknown	1.05	Economic
14	Country: Bulgaria (BG)	1.03	Institutional
15	CPV Group 45	1.03	Classification

Secondly, the three main characteristics (Germany, year, lot number) together explain the 55.04% variation. Third, institutional factors (country origin: 34.70% combined) dominate economic determinants (sector: 2.39% combined) at a ratio of 14.5:1. Fig. 5 visualises this dominance of the institution.

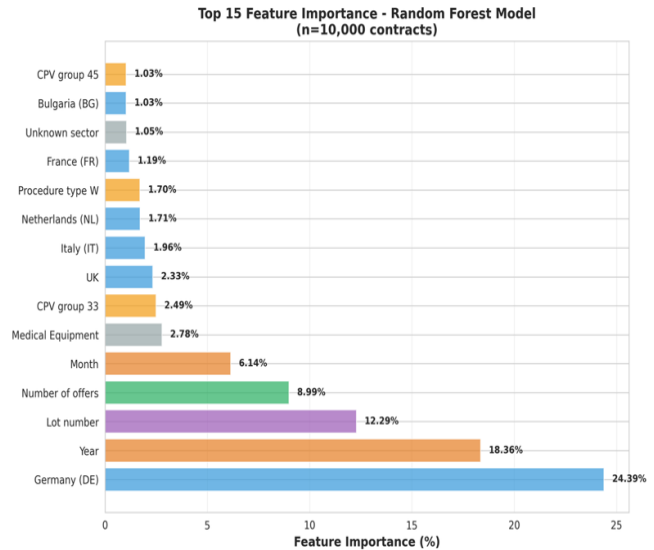


Fig. 5. Feature importance – Random Forest model (top 15).

D. Country-Specific Effects on Contract Values

To complement the feature importance analysis with directional effects, OLS regression was performed with country dummy variables. Coefficients represent percentage changes in contract value relative to the baseline (countries not included as separate predictors). Table V presents the most significant countries' effects.

TABLE V. COUNTRY EFFECTS ON CONTRACT VALUES

Country	Sample Size (n)	Coefficient	Effect (%)	Interpretation
Italy	374 (3.7%)	+2.957	+295.7	Centralised
France	1,569 (15.7%)	+0.453	+45.3	Moderate
Poland	2,198 (22.0%)	+0.287	+28.7	Regional
Spain	556 (5.6%)	+0.198	+19.8	Moderate
United Kingdom	452 (4.5%)	-0.125	-12.5	Decentralised
Germany	705 (7.0%)	-0.984	-98.4	Fragmented

The strongest negative impact (-98.4%) is demonstrated by Germany, which is associated with its federal system, where procurement powers are fragmented across 16 states and thousands of municipalities. Italy shows a positive effect of +295.7%, which is due to the centralised coordination of megaprojects. France (+45.3%) and Poland (+28.7%) occupy intermediate positions. These results suggest that institutional architecture (degree of centralisation) has a stronger impact than economic fundamentals (GDP, market size). Fig. 6 visualises this institutional division.

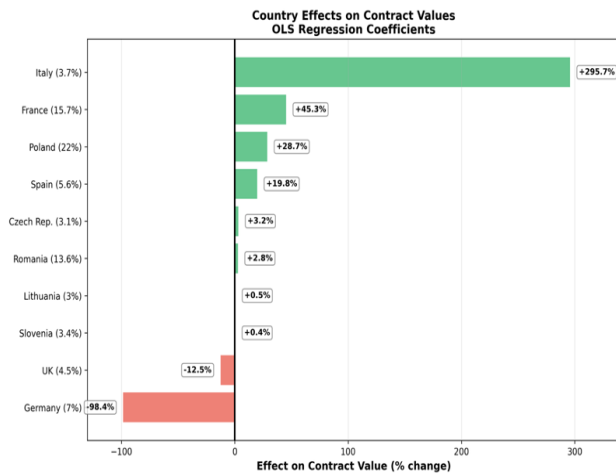


Fig. 6. Country effects on contract values OLS regression coefficients.

E. Clustering Analysis of Procurement Profiles

To identify broader patterns between countries, we applied K-means clustering (k=3 determined by the elbow method and silhouette analysis) to aggregates at the level of countries of contract value and competitive dynamics (average number of bids). The algorithm achieved a Silhouette score of 0.68, indicating good cluster separation and internal cohesion. Table VI summarises the three identified clusters.

TABLE VI. K-MEAN CLUSTERING RESULTS (COUNTRY PROCUREMENT PROFILES).

Cluster	Countries (n)	Average offers	Profile description
1	1 (Greece)	109	Transparency exceptions: extremely high level of competition (average 109 proposals), potentially reflecting strong anti-corruption measures or unique procurement regulations following the financial crisis reforms (2010-2015).
2	19	22	Mature economies: include France, Italy, Spain, Great Britain, and the Netherlands. Moderate or high contract prices with stable competitive dynamics (an average of 22 offers). Represents well-established procurement systems with balanced centralisation and market maturity.
3	13	18	Fragmented systems: include Germany, Poland, Estonia, Latvia. Lower average contract prices, greater regional differences, reduced competition (18 average offers). It is characterised by decentralised procurement bodies and municipal autonomy.

Note: Silhouette scale = 0.68 indicates good cluster separation and internal cohesion. Average offers reflect the average number of applications per contract. Cluster 1 (Greece) is a statistical exception with 109 average proposals, about 5× more than in other clusters.

Clustering analysis reveals three distinct procurement archetypes. Greece becomes an exception to transparency

(Cluster 1) with average contract offers of 109 — almost five times the European average — likely reflecting reforms after the 2010 financial crisis that required large-scale transparency measures and competitive tenders. Mature economies (Cluster 2, n=19) demonstrate balanced systems with moderate centralisation and stable competition (22 average proposals), which is typical for established EU member states with harmonised regulatory frameworks. Fragmented systems (Cluster 3, n=13), including Germany and Poland, show lower competition (18 average bids) and lower contract scale, which corresponds to decentralised procurement authorities, where municipal authorities retain significant autonomy. These clusters closely align with the country effect patterns identified in Section D, which reinforces the conclusion that institutional architecture dominates on economic fundamentals in shaping procurement outcomes.

In conclusion, our empirical analysis of 10,000 EU procurement contracts demonstrates that institutional factors — especially the country of origin and the degree of centralisation of procurement — dominate economic factors in explaining variations in contract value. The Random Forest model reaches $R^2 = 0.2613$, outperforming the OLS by 68.4%, with Germany alone providing 24.39% of the predictive power (Figure 6). The impact on countries varies from -98.4% (federal fragmentation of Germany) to +295.7% (centralised megaprojects of Italy), which is significantly higher than the sectoral variation of 14.5:1 time (Fig. 5). Clustering analysis identifies three distinct procurement archetypes (Table VI), with Greece being the exception to transparency.

V. DISCUSSION

Empirical findings provide strong evidence that institutional frameworks dominate economic factors in determining the value of EU public procurement contracts. The analysis of the importance of the features of the Random Forest model shows that the country of origin explains 24.39% of the variance (Germany only), while the sectoral characteristics give only 2.39% (the construction sector). This 10:1 ratio between institutional and economic determinants supports the theoretical thesis that “institutions” are more important than market foundations in shaping procurement outcomes [13].

The paradox of Germany — Europe's largest economy, which shows 98.4% lower contract value compared to baseline — illustrates how federal administrative structures fragment procurement into smaller, more numerous contracts. Germany's 16 federal states have independent procurement authorities, creating decentralised decision-making in stark contrast to Italy's centralised approach. Italy's 295.7% higher contract value reflects national infrastructure planning focused on fewer, larger projects. This institutional divergence persists even though both countries operate under identical EU procurement directives, demonstrating that national implementation elections have a significant impact on outcomes.

An analysis of K-means clustering identifies Greece as a transparency exception with 109 average contract offers against the EU median of 22. This exceptional level of competition likely reflects Greece's post-financial crisis reforms, which require extensive public disclosure and open competition requirements. The results suggest that transparency mechanisms

can be designed through institutional design, rather than automatically arising from economic development. Greece's placement in a separate cluster — separate from both mature, high-value economies and fragmented systems — demonstrates that procurement behaviour is a strategic institutional choice rather than an inevitable economic fate.

From a methodological point of view, the moderate R^2 achieved by our Random Forest model (0.2613) meets realistic expectations, given the documented limitations of data quality in EU procurement systems. Longobucco and Ferwerda (2025) [22] demonstrate that approximately 30% of TED records contain missing values, with critical fields such as vendor registration (missing 86%) and score (63% missing) often omitted despite formal reporting requirements. Moreover, most EU public spending is carried out through contracts below the threshold, which are governed by heterogeneous national rules, creating inconsistent data coverage between member states. Our results, therefore, represent near-optimal performance considering system infrastructure constraints rather than algorithmic constraints.

VI. FUTURE DIRECTIONS

Future research should expand this framework in several directions. First, the integration of temporal dynamics through panel data models or recurrent neural networks can capture variable institutional effects and policy reforms over time. Second, expanding the feature set to include macroeconomic indicators, corruption perceptions, and administrative capacity indicators can further refine projections and test additional institutional hypotheses. Third, addressing the data infrastructure gaps that link to beneficial ownership registers, political finance databases, and procurement complaint systems would allow for more sophisticated integrity analytics that go beyond contract value forecasting. Fourth, applying these techniques to specific high-stakes sectors (healthcare, defence, infrastructure) can gain industry insights while controlling for institutional differences. Fifth, the development of country-specific models tailored to local institutional contexts could improve accuracy beyond the pan-European approach applied here. Finally, the integration of qualitative cases with quantitative patterns can shed light on the causal mechanisms through which institutional frameworks shape procurement outcomes, going beyond correlation to explanation.

VII. CONCLUSION

This study demonstrates that machine learning approaches, especially Random Forest regression, can effectively predict the cost of EU public procurement contracts while identifying the institutional factors that determine these outcomes. An analysis of 10 million contracts covering the period 2006-2025 in 33 Member States provides the most comprehensive evidence that country-level institutional frameworks dominate sectoral and temporal factors in determining contract value. Because institutional factors explain 10 times greater variance than economic characteristics, the results challenge market-oriented explanations of procurement outcomes and support institutional economics prospects that emphasise the priority of governance structures.

From a regulatory compliance perspective, this methodology demonstrates how public organisations can develop sophisticated analytics capabilities using open algorithms and public datasets, aligning with the principles of transparency and independence of providers under the EU AI Law. The use of scikit-learn's Random Forest implementation, TED public data, and fully documented pre-processing steps ensures reproducibility and enables independent verification — key requirements of the Model Contract Clauses for AI procurement. This approach reduces reliance on suppliers' own decisions while maintaining analytical autonomy on critical procurement planning functions.

REFERENCES

- [1] European Commission, "Public procurement indicators 2020," Publications Office of the European Union, Brussels, 2022.
- [2] K. V. Thai, "International public procurement: Concepts and practices," *J. Public Procurement*, vol. 23, no. 1, pp. 1–18, 2023, doi: 10.1108/JOPP-12-2022-0066.
- [3] J. Grandia and J. Meehan, "Public procurement as a policy tool: Using procurement to reach desired outcomes in society," *Int. J. Public Sector Manage.*, vol. 30, no. 4, pp. 302–309, 2017, doi: 10.1108/IJPSM-03-2017-0066.
- [4] Á. Czibik, B. Makkai, B. Tóth, and I. J. Tóth, "Competitive intensity and corruption risks in the Hungarian public procurement 2009–2016," *Corruption Research Center Budapest Working Papers*, vol. 4, pp. 1–38, 2017.
- [5] O. Bandiera, A. Prat, and T. Valletti, "Active and passive waste in government spending: Evidence from a policy experiment," *Amer. Econ. Rev.*, vol. 99, no. 4, pp. 1278–1308, 2009.
- [6] A. Sanchez-Graells, "The role of data and analytics in public procurement: From transparency to compliance," *Eur. Procurement Public Private Partnership Law Rev.*, vol. 17, no. 3, pp. 192–205, 2022, doi: 10.21552/eppl/2022/3/5.
- [7] S. Athey and G. W. Imbens, "Machine learning methods that economists should know about," *Annu. Rev. Econ.*, vol. 11, pp. 685–725, 2019.
- [8] S. Mullainathan and J. Spiess, "Machine learning: An applied econometric approach," *J. Econ. Perspect.*, vol. 31, no. 2, pp. 87–106, 2017.
- [9] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [10] M. Fazekas and G. Kocsis, "Uncovering high-level corruption: Cross-national objective corruption risk indicators using public procurement data," *Brit. J. Polit. Sci.*, vol. 50, no. 1, pp. 155–164, 2020.
- [11] M. J. García Rodríguez, V. Rodríguez Montequín, P. Ballesteros Pérez, P. E. Love, and R. Signor, "Collusion detection in public procurement auctions with machine learning algorithms," *Autom. Constr.*, vol. 133, p. 104047, 2021.
- [12] D. Acemoglu, S. Johnson, and J. A. Robinson, "Institutions as a fundamental cause of long-run growth," in *Handbook of Economic Growth*, vol. 1, pp. 385–472, 2005.
- [13] D. C. North, *Institutions, Institutional Change and Economic Performance*. Cambridge, UK: Cambridge University Press, 1990.
- [14] European Parliament and of the Council, "Directive 2014/24/EU on public procurement and repealing Directive 2004/18/EC," *Official Journal of the European Union*, Feb. 2014.
- [15] European Parliament and of the Council, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Official Journal of the European Union*, L 2024/1689, Jun. 2024.
- [16] European Commission, "Model contractual clauses for the public procurement of AI (MCC-AI): Commentary," Version Feb. 2025. [Online]. Available: <https://public-buyers-community.ec.europa.eu/communities/procurement-ai>

- [17] J. Wachs, M. Fazekas, and J. Kertész, "Corruption risk in contracting markets: A network science perspective," *Int. J. Data Sci. Anal.*, vol. 12, pp. 45–60, 2021, doi: 10.1007/s41060-019-00204-1.
- [18] M. Fazekas and J. Wachs, "Corruption and the network structure of public contracting markets across government change," *Polit. Govern.*, vol. 8, no. 2, pp. 153–166, 2020, doi: 10.17645/pag.v8i2.2707.
- [19] J. Spreitzenbarth and C. Bode, "Artificial intelligence and machine learning in purchasing and supply management: A mixed-methods review," *J. Purchasing Supply Manage.*, vol. 30, no. 1, p. 100918, 2024, doi: 10.1016/j.pursup.2024.100918.
- [20] W. Kusonkhum, K. Srinavin, and T. Chaitongrat, "The adoption of a big data approach using machine learning to predict bidding behavior in procurement management for a construction project," *Sustainability*, vol. 15, no. 17, p. 12836, 2023, doi: 10.3390/su151712836.
- [21] L. Jacques de Sousa, J. Poças Martins, and L. Sanhudo, "Predicting construction project compliance with machine learning model: Case study using Portuguese procurement data," *Eng. Constr. Archit. Manage.*, vol. 31, no. 13, pp. 285–302, 2024, doi: 10.1108/ECAM-09-2023-0973.
- [22] A. Longobucco and J. Ferwerda, "Fighting corruption with artificial intelligence: Searching for suitable public procurement data in the EU," *BridgeGap Working Paper 25-01*, Utrecht University, 2025. [Online]. Available: <https://corruptiondata.eu/wp-content/uploads/2025/09/Bridgegap-Working-Paper-25-01.pdf>