A Pipeline for Integrating and Analyzing Public Business Data in Albania Using Web Scraping and Machine Learning

Markela Muca, Klodiana Bani, Amarildo Alla
Department of Applied Mathematics-Faculty of Natural Sciences, University of Tirana, Albania

***Abstract*—The increasing availability of fragmented public business and procurement data creates opportunities for company-level empirical analysis, but limited evidence exists on how integrated procurement and administrative records can be used to characterize heterogeneous economic operators in Albania. This study investigates the analytical value of integrating public procurement records from the Public Procurement Agency (APP) with administrative business information from the National Business Center (QKB) through the unique NIPT identifier. The resulting company-level dataset is analyzed to examine whether procurement-derived characteristics contain systematic information for distinguishing heterogeneous company profiles and reproducing heuristic risk-oriented classifications. The empirical framework combines data normalization, entity-level integration, feature engineering, heuristic-label construction, supervised classification, Principal Component Analysis (PCA), K-Means clustering, anomaly detection, feature-importance analysis, and a financial-enrichment sensitivity experiment. In the 18-feature reduced experiment, HistGradientBoosting achieved the strongest mean repeated-cross-validation performance, with an F1-score of 0.8590, ROC AUC of 0.9683, and Average Precision (AP) of 0.9399. The analysis further identifies distinct company profiles characterized by differences in procurement intensity, contract value, and economic exposure, including small groups with exceptionally high activity and atypical characteristics. Financial enrichment did not produce a substantial improvement in classification performance, indicating that procurement-derived variables provide the dominant analytical signal in the current dataset. Across 999 deterministic label permutations, the mean ROC AUC was 0.4999, and the mean AP was 0.2700; the one-sided empirical p-value was 0.001 for both metrics. This diagnostic does not prove the absence of leakage, overfitting, or residual proxy dependence. The findings provide empirical evidence that company-level integration of procurement and administrative data can reveal heterogeneous procurement profiles and atypical patterns relevant to procurement monitoring, public-sector analytical capacity, and data-driven governance. Because the target labels are heuristic and are not independently validated against official risk outcomes, the results are interpreted as exploratory evidence for company profiling and risk-oriented analysis rather than as formal assessments of business risk, misconduct, or performance.**

Keywords*—Web Scraping; machine learning; data integration; public data analysis; public procurement; benchmarking; Albania

I. Introduction

The increasing availability of public data has expanded opportunities for empirical analysis of businesses and for supporting decision-making in public administration, scientific research, and the private sector. Administrative registries and public procurement data contain valuable information on company characteristics, economic activity, and relationships with public institutions. However, these data are frequently distributed across heterogeneous sources with different structures, standards, identifiers, and levels of completeness, making integrated company-level analysis challenging [1], [2], [23], [26].

In Albania, two important sources of public business information are the Public Procurement Agency (APP) and the National Business Center (QKB). APP provides information on public procurement procedures and participating or winning economic operators, whereas QKB provides administrative and identification information on registered economic entities [27], [28]. Although each source provides valuable information independently, their integration creates an opportunity to examine economic operators across multiple procurement activities while simultaneously considering their administrative characteristics.

Integrating these sources presents methodological challenges related to heterogeneous formats, inconsistent representations, entity identification, missing information, and data quality. Consequently, reliable analysis requires data collection, normalization, quality control, identifier harmonization, and feature construction [2], [23], [26]. The present study does not consider these procedures to be the primary scientific contribution. Rather, they provide the empirical foundation for investigating what additional information about companies becomes observable when fragmented procurement and administrative records are integrated at the company level.

Existing research has examined procurement transparency and data-driven governance [1], [21], [25], while other studies have focused on corruption-risk indicators [6], [10], [11] and machine learning for procurement monitoring [7], [8], [33]. Procurement data have been used to construct objective corruption-risk indicators and to identify procurement characteristics associated with potential irregularities [6], [10], [11]. Machine Learning has also been applied to procurement complaints, anomalies, bidder classification, and oversight prioritization [7], [8], [33].

However, much of this literature operates at the procurement-procedure or contract level, focusing on individual tenders, contracting authorities, complaints, competition

indicators, or procurement-risk signals. Comparatively less empirical attention has been given to the economic operator as the analytical unit across multiple procurement procedures. This distinction is important because aggregating procurement activity at the company level makes it possible to examine heterogeneous economic operators according to their procurement intensity, contract exposure, administrative characteristics, and atypical activity patterns.

Evidence from Albania is particularly relevant. Martiri and Kapedani [9] examined competition in Albanian public procurement of services using Public Procurement Agency data for 2010–2022 and demonstrated the analytical value of procurement indicators for understanding competition and accountability. However, this research focuses primarily on procurement competition rather than on integrated company-level profiling.

The research gap addressed in this study therefore concerns the limited empirical evidence on the structure of company-level procurement profiles obtained by integrating public procurement records with administrative business-registry information, particularly in the Albanian context. The gap has three dimensions. First, existing procurement research predominantly evaluates individual procedures or procurement outcomes rather than the accumulated activity of economic operators across multiple procedures [6], [7], [8], [9], [10], [11], [33]. Second, limited evidence is available on whether integrated procurement and administrative information reveal distinguishable groups of companies with different levels of economic activity and exposure. Third, when heuristic labels are constructed from procurement signals, it remains necessary to determine whether meaningful discriminatory information persists after variables directly contributing to label construction are removed.

This study addresses these gaps by examining a company-level dataset obtained by integrating APP procurement records with QKB administrative information through the NIPT identifier [27], [28]. The analysis investigates whether procurement-derived characteristics provide systematic information for distinguishing heterogeneous company profiles and for reproducing heuristic risk-oriented classifications under a reduced-feature design. The study also examines whether available financial information contributes additional discriminatory information beyond procurement-derived characteristics [29].

The empirical analysis combines supervised and unsupervised approaches. Supervised classification is used to evaluate the discriminatory information contained in the engineered company-level features, whereas Principal Component Analysis (PCA), K-Means clustering, and anomaly detection are used to identify and characterize heterogeneous company profiles independently of the classification task [16], [18], [19], [31], [32]. Feature-importance analysis is subsequently used to examine which procurement-related variables contribute most strongly to the classification models [12]. A financial-enrichment experiment further evaluates whether the addition of available financial indicators provides information beyond procurement-derived characteristics. An important methodological issue concerns the construction of the target labels. Because official company-level risk outcomes are not available in the dataset, two heuristic labels are constructed from procurement-related signals. The first experiment evaluates the ability of the models to reproduce the weaker heuristic label using the full feature set. The second experiment provides a more conservative sensitivity analysis by removing features that directly contribute to the construction of the stricter heuristic label and evaluating the models using the remaining 18 features. Consequently, the classification results are not interpreted as evidence of formally validated business risk, corruption, misconduct, or business failure. Instead, they indicate the extent to which the available company-level characteristics contain information associated with the heuristic profiles defined for this study.

The study makes four substantive contributions. First, it provides empirical evidence on the analytical value of integrating fragmented public procurement and administrative business data at the company level in Albania. Second, it identifies heterogeneous company profiles based on procurement intensity, contract value, and administrative characteristics, including small groups exhibiting exceptionally high levels of activity and atypical economic exposure. Third, it evaluates the persistence of discriminatory information after reducing direct feature-label overlap. Fourth, it examines the incremental value of financial enrichment.

These contributions distinguish the study from work that primarily develops generic data-processing or Machine Learning pipelines. The computational procedures are used as analytical instruments, while the principal contribution lies in the empirical evidence generated from the integrated APP–QKB dataset and in the company-level perspective that this integration enables. The findings are therefore positioned as exploratory evidence relevant to procurement monitoring, public administration, and data-driven governance rather than as an automated institutional risk-assessment system [8], [21], [25].

The remainder of the study is organized as follows. Section II presents the related work on public procurement analytics, Machine Learning for procurement monitoring, administrative data integration, company-level profiling, and research in Albania and comparable contexts. Section III describes the proposed methodology, including the data sources, collection and integration procedures, feature engineering, heuristic-label construction, and experimental design. Section IV presents the classification results, PCA and clustering analysis, anomaly detection, feature-importance analysis, and financial-enrichment experiment. Section V discusses the findings and their implications for procurement governance, public administration, and digital government. Section VI presents the conclusions and limitations, while Section VII outlines directions for future research.

II. Related Work

This section reviews empirical research relevant to the analytical use of public procurement data, Machine Learning for procurement monitoring, administrative data integration, and company-level profiling. Particular attention is given to evidence from Albania and comparable European and post-

transition contexts, thereby positioning the present study within its specific institutional and application domain.

A. Public Procurement Data and Transparency

Public procurement data provide an important empirical basis for analyzing competition, transparency, efficiency, accountability, and potential irregularities in government contracting. The availability of open government data has increasingly created opportunities to move from descriptive transparency toward systematic empirical analysis of public-sector activity [1], [21], [25]. However, the analytical value of these data depends strongly on their quality, completeness, standardization, and consistency. Research on data and information quality emphasizes that inconsistencies, missing information, heterogeneous structures, and inadequate data preparation can substantially affect the reliability of subsequent analytical processes [2], [23], [26].

Empirical studies have demonstrated that procurement data can be used to construct risk indicators and identify procurement procedures that warrant additional institutional scrutiny. Lisciandra, Milani, and Millemaci [6] developed a corruption risk indicator using Italian public works data, while Fazekas and colleagues [10], [11] demonstrated the usefulness of objective procurement indicators for identifying patterns associated with corruption risk.

The present study builds on this literature by examining whether procurement information can be aggregated across multiple procedures to characterize economic operators and identify heterogeneous company-level profiles.

B. Machine Learning for Procurement Monitoring

Machine Learning has increasingly been used for classification, anomaly detection, risk assessment, and decision support in public procurement [3], [7], [8], [14], [15]. Nai et al. [7] combined public procurement data with administrative justice records in Italy and developed Machine Learning models to identify procurement procedures associated with complaints and disputes. Salazar, Pérez, and Gallego [8] developed VigIA, combining Machine Learning models with risk indices to support procurement oversight and prioritization.

Recent work in Romania has also applied Machine Learning to a large public procurement dataset for bidder classification, illustrating the potential of large-scale procurement data for automated analytical assessment [33].

An important methodological issue in this literature concerns the relationship between heuristic target construction and model predictors. When a target variable is derived from procurement characteristics that are subsequently used as model inputs, the resulting performance may partly reflect label-feature dependency. The present study therefore distinguishes between a full-feature experiment and a reduced-feature experiment designed to examine whether discriminative information remains after removing features directly involved in constructing the stricter heuristic label. Model evaluation follows established principles for classification and imbalanced data, using cross-validation [17], ROC AUC [5], and Precision-Recall measures [24].

C. Administrative Data Integration and Company-Level Business Profiling

Administrative data are frequently collected and maintained by different public institutions for distinct administrative purposes. Their integration can create analytical opportunities that are not available when each source is considered independently [1], [2], [23], [26].

In the context of public procurement, procurement records describe the interaction between economic operators and contracting authorities, whereas business registries provide additional information about the economic operators themselves [27], [28].

Existing procurement studies have primarily examined tenders, contracts, contracting authorities, competition indicators, complaints, and procurement-specific risk signals [6], [7], [8], [9], [10], [11], [33]. Although this literature provides substantial evidence at the procurement-procedure level, comparatively less attention has been given to the economic operator as an analytical unit across multiple procurement activities.

A company-level perspective provides a complementary analytical dimension. Instead of evaluating an individual tender in isolation, an integrated dataset can characterize an economic operator according to procurement participation, contract values, activity intensity, administrative characteristics, and atypical patterns.

For the present study, the NIPT identifier provides the linkage mechanism between procurement and business-registry information [27], [28]. This linkage enables procurement observations to be aggregated at the company level and provides an empirical basis for examining heterogeneity among economic operators participating in public procurement in Albania.

D. Public Procurement Research in Albania

Research on Albanian public procurement has examined competition, savings, transparency, accountability, and institutional performance. Martiri and Kapedani [9], using Public Procurement Agency data covering 2010–2022, examined competition in Albania's public procurement of services through indicators including bids per tender, funding allocations, contract numbers, disqualified bids, and single- versus multiple-bid contracts. Their findings demonstrate the relevance of procurement data for evaluating the functioning of the Albanian procurement market.

These findings establish the empirical relevance of procurement analysis in Albania, but they primarily examine procurement competition and institutional aspects rather than integrated company-level profiles. Evidence from Italy, Colombia, and Romania further demonstrates the potential of combining procurement data with Machine Learning and administrative information [7], [8], [33]. However, differences in institutional structures, registry systems, procurement practices, and data availability mean that findings from other countries cannot simply be transferred to Albania without empirical examination of the Albanian data environment.

E. Research Gap

Existing research has established the value of public procurement data, Machine Learning, risk indicators, anomaly detection, and data-driven monitoring for public-sector decision support [6], [7], [8], [9], [10], [11], [18], [21], [25], [33]. Research on data quality and administrative data integration further emphasizes the importance of preparing and linking heterogeneous public datasets before conducting empirical analysis [2], [23], [26].

The research gap addressed in this study is the limited empirical analysis of economic operators through the integration of public procurement records and administrative business-registry information at the company level, particularly in the Albanian context.

The present study addresses this gap by integrating APP procurement records with QKB administrative business information through the company NIPT identifier [27], [28]. The resulting company-level dataset is used to investigate company profiles, the discriminatory information retained after reducing direct label-feature overlap, the characteristics contributing to company differentiation, and the incremental value of available financial information [29].

The contribution therefore lies primarily in the empirical characterization of Albanian economic operators enabled by the integration of APP and QKB data, rather than in the development of a new Machine Learning algorithm.

III. Proposed Methodology

A. Research Design and Analytical Framework

The study adopts an empirical, company-level analytical design to investigate the analytical value of integrating public procurement records with administrative business information. The objective is not to develop a new Machine Learning algorithm, but to examine whether the integration of fragmented public data enables the identification and characterization of heterogeneous economic operators participating in public procurement.

The methodological framework consists of six interconnected stages: 1) data collection, 2) data normalization and quality control, 3) company-level data integration, 4) feature engineering and heuristic-label construction, 5) supervised and unsupervised analysis, and 6) robustness and sensitivity evaluation. The overall process transforms heterogeneous record-level public information into a standardized company-level analytical dataset.

The analytical unit is the company, identified through the NIPT/NUIS identifier. This choice distinguishes the present study from procurement-level approaches that analyze individual tenders or contracts. By aggregating procurement observations at the company level, the analysis captures the accumulated procurement activity of economic operators over the observation period.

The methodological framework combines supervised classification with complementary unsupervised techniques. Classification models are used to evaluate whether company-level features contain discriminatory information associated with the heuristic labels. PCA and K-Means are used independently of the classification task to investigate the underlying structure and heterogeneity of the companies. Isolation Forest is subsequently used to identify observations that deviate substantially from the dominant feature-space structure [16], [18], [19].

The study therefore treats machine learning as an analytical instrument for empirical investigation rather than as the principal methodological contribution.

B. Data Sources

The experiment uses two primary administrative sources and one supplementary financial source.

1) The Public Procurement Agency (APP) provides publicly accessible information concerning public procurement procedures and economic operators participating in or winning procurement activities. Historical procurement records covering the period 2006–2025 were collected from the APP public data environment [27].

The collected procurement information includes variables describing procurement procedures, contracting authorities, procedure status, allocated budgets, winning economic operators, and contract values. These records constitute the primary source for measuring the procurement activity of companies.

During the data collection process, 10,341 unique companies that had won at least one public procurement contract were identified.

2) National Business Center (QKB): The National Business Center provides administrative information concerning registered economic entities in Albania [28]. The QKB data were used to enrich procurement observations with company-level administrative characteristics.

A total of 59,266 documents were processed. Relevant attributes included legal form, current business status, location, incorporation information, and company age.

The QKB data therefore provide contextual information about the economic operators identified in the procurement records and allow the analysis to move beyond procurement activity alone.

3) Open Corporates Albania was used as a supplementary source for the financial-enrichment experiment [29]. A total of 40,034 historical financial records were collected.

Financial variables included information on annual revenue and net profit for companies for which such information was available. Because financial coverage was substantially lower than the coverage of procurement and administrative data, financial information was not included in the primary classification dataset. Instead, it was incorporated in a separate sensitivity experiment to assess whether financial variables provide additional discriminatory information beyond procurement-derived characteristics.

C. Data Collection, Normalization, and Quality Control

The data collection stage employed a structured web-scraping process to retrieve publicly available procurement and business information from the relevant sources [20]. During the collection process, Albiz Collector structurally stores process metadata (requested URLs, content hashes, and HTTP status codes), ensuring the provenance and reproducibility of the pipeline.

In total, these sources enabled the construction of an integrated company-level dataset, which served as the foundation for analytical feature engineering and the application of machine learning methods for classification, profiling, and anomaly detection.

Data integration was achieved through the exact matching of the unique Taxpayer Identification Number (NIPT/NUIS), following the standardization of its format. Out of the 10,341 winning companies identified in the APP data, 10,266 companies were successfully matched with the QKB registry.

The financial source covered a smaller subset. Financial information was available for 3,159 companies, corresponding to 30.8% of the 10,266 companies in the integrated APP–QKB dataset.

D. Integration Pipeline and Feature Engineering

Based on this integrated dataset, 30 initial features were constructed, describing public procurement activity, company characteristics, and the ratios between contract values and allocated budgets**.** These features form the basis for exploratory analysis and the development of machine learning models.

To transform the aggregated procurement data into standardized, company-level analytical indicators, two analytical indicators were constructed to summarize procurement activity.

1) Business Performance Indicator (BPI): This procurement-activity proxy combines five nonnegative company-level components after maximum-normalized logarithmic transformation:

$$BPP_i = 100 * \sum_{k=1}^{m} w_k \frac{\ln(1+x_{ik})}{\ln(1+M_k)},\ \mathrm{M_k} > 0$$

where, x_{ik} represents the quantitative component k of the activity of company i (such as the number of won tenders or the total value) and w_k is the specific weight allocated to each component. M_k is its maximum in the 10,266-company analytical dataset, and the fixed weights are 0.30 for active procurement count, 0.30 for active total winner value, 0.20 for distinct contracting-authority count, 0.10 for active-year span, and 0.10 for rows with a winner value. These weights define an analyst-specified procurement-activity proxy; they were not estimated from independent business outcomes. The components, weights, and operational rationales are summarized in Table I.

2) The Sure Winner Ratio (SWR): To measure the efficiency or deviation of winning bids relative to the allocated budgets, without causing division-by-zero errors, the function is applied:

$$SWR_{ij} = \frac{V_{ij}^F}{V_{ij}^B},\ V_{ij}^B > 0$$

where, V_{ij}^F is the winning value of the economic operator, and V_{ij}^B is the initial budget allocated for that procedure.

The Open Contracting Data Standard (OCDS) was used as a conceptual reference for procurement terminology and the organization of procurement information [22]. It was not used as the source of the mathematical formulations above.

TABLE I. COMPONENTS AND FIXED WEIGHTS OF THE PROCUREMENT-ACTIVITY PROXY

Component	Weight	Rationale
Active procurement count	0.30	Overall procurement activity
Active total winner value	0.30	Economic exposure
Distinct contracting-authority count	0.20	Breadth of institutional participation
Active-year span	0.10	Continuity of procurement activity
Rows with a winner value	0.10	Availability of recorded award values

E. Heuristic Risk-Label Construction

Due to the lack of official risk labels, two heuristic labels were constructed (weak_risk_label and strict_weak_risk_label) based on public procurement activity signals. The heuristic Risk Function (y_i) determines whether a company i is assigned to Class 1 or Class 0 according to the number of predefined procurement-related signals triggered by the company.

The general formulation is:

$$C_i = \sum_{k=1}^{q} I_{ki}, \qquad y_i = \begin{cases} 1, & \text{if } C_i \geq \tau \\ 0, & \text{if } C_i < \tau \end{cases}$$

The distribution of companies across the two classes for both heuristic labels is presented in Table II. For the *weak_risk_label*, 6,889 companies (67.1%) belong to Class 0 and 3,377 companies (32.9%) to Class 1. For the more restrictive *strict_weak_risk_label*, Class 0 contains 7,521 companies (73.2%), while Class 1 contains 2,745 companies (26.8%). Thus, both label variants exhibit a moderate class imbalance, with the imbalance being more pronounced for the strict label.

TABLE II. HEURISTIC LABEL DISTRIBUTION AND DATASET CLASS STRUCTURE

Variable	Class 0	Class 1	Total number of companies
weak_risk_label	6,889 (67.1%)	3,377 (32.9%)	10,266
strict_weak_risk_label	7,521 (73.2%)	2,745 (26.8%)	10,266

Two experimental settings were considered to evaluate the classification models while examining the potential dependence between the predictors and the heuristic labels.

The first, Full-Feature, Weak-Label Replication, used the complete feature set to reproduce the *weak_risk_label*. Because some of the features were also involved in constructing the heuristic label, the results of this experiment primarily indicate

how well the models reproduce the structure embedded in the labeling rule. The second, Reduced-Feature, Strict-Label, excluded the features that directly contributed to the construction of the *strict_weak_risk_label*, leaving 18 features for model training. This second setting was introduced as a sensitivity analysis to reduce direct label-feature dependency and to examine whether predictive information remained after the most immediate overlap between the predictors and the heuristic target had been removed.

To assess whether the observed classification performance could arise from random association between the retained features and the heuristic target, a shuffled-label permutation diagnostic was conducted using the 18-feature reduced dataset and HistGradientBoosting under the same repeated 5-fold cross-validation structure. An initial run with 10 permutations was used as a preliminary diagnostic to verify the permutation procedure and null behavior. Because 10 permutations provide a coarse empirical p-value resolution, the analysis was subsequently repeated with 999 deterministic permutations, which was treated as the final diagnostic. For each permutation, the target labels were randomly reassigned while the feature matrix and cross-validation structure were retained unchanged. The empirical one-sided p-value was calculated as $p = (r + 1)/(B + 1)$, where r denotes the number of permutations whose performance equaled or exceeded the observed value and B is the number of permutations.

Six classification algorithms were evaluated: HistGradientBoosting, Random Forest, Gradient Boosting, Extra Trees, KNN, and Logistic Regression. These models were selected to provide a comparison between nonlinear tree-based ensemble approaches and simpler parametric or distance-based methods. Random Forest follows the ensemble framework introduced by Breiman [4], while Gradient Boosting is based on the framework proposed by Friedman [13]. Extra Trees follows the extremely randomized tree approach proposed by Geurts, Ernst, and Wehenkel [30]. General machine learning principles relevant to the classification experiments were also considered [3], [14], [15]. This combination of models allows the analysis to assess whether the nonlinear relationships and interactions present in the procurement-derived features provide additional discriminatory information compared with simpler modeling approaches.

Model performance was evaluated using repeated 5-fold cross-validation with three repetitions. In each repetition, the dataset was divided into five folds. Four folds were used for model training and one fold for validation, with the process repeated until each fold had served as the validation set. The procedure was then repeated three times to reduce the influence of a particular partition of the observations. Cross-validation provides a more robust estimate of model performance than relying on a single train-test partition [17].

The moderate class imbalance shown in Table II was taken into account when evaluating the classification results. Accuracy was therefore not treated as a sufficient standalone measure, since a model can obtain relatively high accuracy by favoring the majority class without performing equally well on the minority class. Model performance was assessed using F1-score, ROC AUC, Precision-Recall AUC (PR AUC), Recall, and Accuracy. F1-score provides a balance between precision and recall, while ROC AUC measures the ability of a classifier to distinguish between the two classes across classification thresholds [5]. PR AUC was also considered because Precision-Recall analysis provides a more informative assessment of classification performance when the classes are not equally represented [24]. Taken together, these measures provide a broader assessment of model performance and reduce the risk of concluding accuracy alone.

IV. Analysis Results and Evaluation of Machine Learning Methods

A. Classification Results

In the first experiment, the models demonstrated a strong capability to reproduce the heuristic label (weak_risk_label). The HistGradientBoosting model achieved the highest performance in this experiment. The potential label-feature dependency was addressed in Experiment 2 (Reduced-Feature), where nine direct label contributors and three close alternate aggregates were excluded; the models were trained on 18 retained source features and evaluated against the strict_weak_risk_label target.

In this configuration, six classification algorithms were compared: HistGradientBoosting, Random Forest, Gradient Boosting, Extra Trees, KNN, and Logistic Regression. Performance was estimated with Repeated Stratified K-Fold cross-validation (five folds, three repetitions, random state 42), resulting in 15 validation folds. The mean performance measures for the Reduced-Feature experiment are presented in Table III.

TABLE III. Mean Repeated-CV Results for the 18-Feature Reduced Experiment

Model	Accuracy	AP	Recall	F1-Score	ROC AUC
Logistic Regression	0.8161	0.7202	0.6674	0.6599	0.8487
Random Forest	0.9168	0.9305	0.8605	0.8469	0.9644
Gradient Boosting	0.9223	0.9297	0.8056	0.8472	0.9634
Extra Trees	0.9092	0.9052	0.7537	0.8160	0.9533
HistGradientBoosting	**0.9283**	**0.9399**	**0.8171**	**0.8590**	**0.9683**
KNN	0.7793	0.5430	0.4137	0.5006	0.7351
HGB shuffled-label null mean. B=10 permutations	0.7254	0.2716	0.0149	0.0281	0.5025

The results of Experiment 2 indicate that HistGradientBoosting achieved the strongest mean performance, with an F1-score (0.8590), ROC AUC (0.9683), and Average Precision (0.9399) among the evaluated models (Table III). Random Forest, Gradient Boosting, and Extra Trees also exhibited high performance with ROC AUC values of 0.9644 and 0.9634, respectively. Extra Trees produced a ROC AUC of 0.9533, whereas Logistic Regression and KNN achieved lower values of 0.8487 and 0.7351. These values describe agreement with a constructed heuristic target and are not validation against independently observed risk outcomes.

To evaluate whether the observed classification performance could be explained by random association between the retained

features and the heuristic target, a shuffled-label permutation diagnostic was performed using the 18-feature dataset and HistGradientBoosting.

Table III(A) shows that the observed mean ROC AUC was 0.9683 and the observed AP was 0.9399. In the final 999-permutation analysis, the null mean ROC AUC was 0.4999 and the null mean AP was 0.2700. No permutation equaled or exceeded the observed performance for either metric. Thus, with r=0 and B=999, the one-sided empirical permutation p-value was 0.001 for both ROC AUC and AP.

TABLE III (A). SHUFFLED-LABEL PERMUTATION DIAGNOSTIC FOR THE 18-FEATURE HISTGRADIENTBOOSTING MODEL

Metric	Observed	Null Mean (B=10)	Empirical p-value (B=10)	Null Mean (B=999)	Empirical p-value (B=999)
ROC AUC	0.9683	0.5025	0.0909	0.4999	0.001
AP	0.9399	0.2716	0.0909	0.27	0.001

The preliminary 10-permutation run produced null means of 0.5025 for ROC AUC and 0.2716 for AP, with a minimum attainable p-value of 0.0909 under the same plus-one estimator.

The results indicate that the observed discrimination is substantially higher than the performance obtained under randomly permuted labels. Thus, the permutation diagnostic provides evidence against a purely random feature-label association. Nevertheless, it does not establish the absence of data leakage, overfitting, or residual proxy dependence resulting from the heuristic construction of the target.

B. Principal Component Analysis (PCA) and Clustering

To explore the dataset structure in an unsupervised manner, Principal Component Analysis (PCA) was applied to the standardized numerical features. PCA enables data dimensionality reduction by transforming the original variables into a new set of components, ordered by the amount of variance they explain [16]. The objective of applying PCA in this study was not to eliminate features for predictive modeling, but rather to identify latent structures and visualize company profiles in a lower-dimensional space.

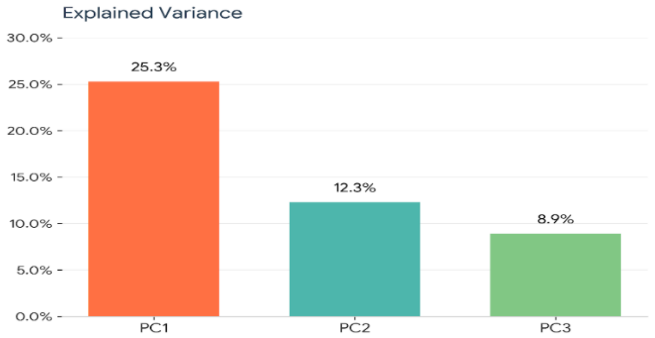

Fig. 1. Explained variance ratio for the first three components, PCA.

As shown in Fig. 1, the first three principal components explained 25.3%, 12.3%, and 8.9% of the total variance, respectively, accounting for approximately 46.5% of the total information in the dataset. The analysis of component loadings revealed that PC1 is dominated by features related to the scale of public procurement activity, such as the number of active contracts and the total contract value. Subsequent components are more closely associated with administrative characteristics and ratios between contract values. This indicates that the primary structure of the dataset is determined by the economic intensity of public activity, whereas administrative characteristics contribute to the secondary differentiation of profiles.

K-Means clustering was applied to the same standardized and one-hot encoded 51-dimensional profile matrix used by the implementation; PCA was used only for visualization. Solutions from k = 2 to k=10 were evaluated to examine the sensitivity of the cluster structure to the selected number of groups.

The clustering diagnostics favored k=2, with a silhouette coefficient of 0.675 and a Calinski-Harabasz index of 1,582.1. In comparison, the k=5 solution produced a substantially lower silhouette coefficient of 0.192 and included a very small cluster containing only 18 companies. Inertia decreased progressively as increased, without indicating a unique elbow point (Fig. 2(a), Fig. 2(b)).

Therefore, the existing k=5 solution is retained as a higher-resolution exploratory representation consistent with the implemented analytical dashboard, rather than as the empirically optimal solution for identifying natural company groupings. The five clusters should consequently be interpreted as exploratory company profiles rather than definitive or naturally separated company types.

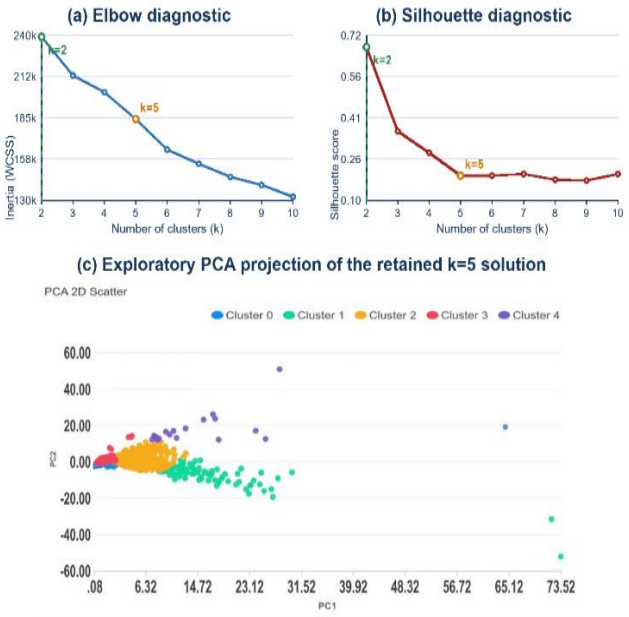

Fig. 2. Cluster-count sensitivity and retained exploratory view: (a) Inertia and (b) Silhouette score for k = 2–10; (c) PCA projection of the existing k = 5 solution.

The resulting five profiles are presented in Table IV. The two largest groups (Cluster 0 and Cluster 3), which represent approximately 82.8% of the companies, are characterized by relatively low public procurement activity. Cluster 2 comprises companies with high contract values and a higher proportion of positive heuristic-label assignments, whereas Cluster 1 and

Cluster 4 represent very small groups of companies with exceptionally high levels of activity and a higher concentration of atypical observations, suggesting the presence of outliers in the dataset.

TABLE IV. KEY CHARACTERISTICS OF THE CLUSTERS IDENTIFIED BY K-MEANS ALGORITHM

Clusters	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Companies	5,294	95	1,658	3,201	18
Share	51.6%	0.9%	16.2%	31.2%	0.2%
Mean performance	30.51	79.30	58.49	27.65	69.91
Mean active count	7.00	920.32	91.43	4.69	233.00
Mean active winner value (M)	9.40M	636.10M	502.02M	10.84M	16,652.14M
Weak label rate	27.8%	100.0%	53.5%	28.6%	33.3%
Strict label rate	21.6%	100.0%	45.0%	23.5%	27.8%
Profile	Low activity	Higher anomaly concentration	High value winners	Low activity	Higher anomaly concentration

Fig. 2(c) illustrates the projection of companies onto the two-dimensional space defined by the first two principal components of the PCA, where the color indicates the cluster membership identified by the K-Means algorithm. The positioning of companies within this space reflects the similarities and differences among their economic and operational profiles, as determined by the features utilized in the analysis. Companies located in proximity to one another exhibit similar structures regarding public procurement activity, performance, and procurement-related indicators, whereas the groups situated further from the center represent more specific profiles or extreme cases.

Cluster 0 (blue) is the largest group, comprising 5,294 companies (51.6% of the dataset). In the PCA space, it appears concentrated near the central zone, indicating a profile similar to the average characteristics of the companies in the dataset. According to Table IV, this cluster is characterized by low public procurement activity (averaging 7 active contracts with a mean value of approximately 9.4 million), low average performance (30.51), and a relatively low level of positive heuristic-label assignments (27.8% under the weak label and 21.6% under the strict label). This suggests that Cluster 0 represents the baseline profile of companies, featuring limited exposure to public contracts and lacking pronounced extreme characteristics.

Cluster 3 (red) represents the second-largest group, comprising 3,201 companies (31.2%). Although it appears in the PCA plot as a distinct group separated from Cluster 0, its characteristics are similar: low public activity (averaging 4.69 active contracts), relatively low contract value (10.84 million), and low performance (27.65). The positive heuristic-label rate is slightly higher compared with Cluster 0 (28.6% under the weak label and 23.5% under the strict label), but the difference is not substantial. Its separate positioning in the PCA space suggests that this cluster differs from Cluster 0 not by the absolute level of activity, but rather by the combination of other features included in the analysis. Consequently, Cluster 0 and Cluster 3 together represent the dominant segment of companies with limited activity, accounting for approximately 82.8% of the dataset.

Cluster 2 (orange) comprises 1,658 companies (16.2%) and presents a significantly different profile from the two primary groups. In Table IV, this cluster is characterized by much higher public procurement activity (averaging 91.43 active contracts), a high mean contract value (502 million), and higher performance (58.49). However, this level of activity is accompanied by a greater prevalence of the procurement signals incorporated into the heuristic labels (53.5% under the weak label and 45.0% under the strict label).

The position of Cluster 2 in the PCA plot is linked to this combination of high economic activity and greater prevalence of heuristic procurement signals, indicating a group of companies of higher economic significance that may consequently warrant more careful analytical monitoring.

Cluster 1 (green) and Cluster 4 (purple) are the smallest and most extreme groups. Cluster 1 includes only 95 companies (0.9%), yet is characterized by exceptionally high levels of activity (averaging 920 active contracts) and a mean contract value exceeding 636 million. All companies in this cluster are assigned to the positive class under both heuristic labels. In the PCA plot, this group appears shifted away from the main body of the dataset, reflecting its extreme and heterogeneous nature. This finding suggests a concentration of cases with very high economic exposure and potentially atypical characteristics.

Cluster 4 (purple) is the smallest group, with only 18 companies (0.2%), but it features the highest mean value of active contracts (16.65 billion). Its isolated positioning in the PCA space aligns with the characteristics shown in the table, indicating companies with a very large scale of activity and a distinct economic profile. Although the risk rate is lower than in Cluster 1 (33.3% under the weak label and 27.8% under the strict label), the extremely high magnitude of financial exposure makes this group highly significant for further analysis and anomaly detection.

Overall, the PCA and K-Means analysis reveals the existence of several distinct company profiles within the Albanian public procurement market. While the vast majority of companies are characterized by limited activity, a smaller subset exhibits significantly higher levels of activity and financial exposure. These extreme profiles serve as a foundation for subsequent anomaly analysis using dedicated outlier detection algorithms, which are presented in the following section.

C. *Anomaly Detection, Feature Importance, and the Financial Enrichment Experiment*

For the identification of profiles with atypical behavior in the feature space, the unsupervised Isolation Forest algorithm was applied. This method is highly suited for high-dimensional datasets, as it identifies observations that are isolated more quickly from the dominant portion of the data and classifies them as potentially anomalous cases [18].

The results demonstrated that the majority of companies lie within the dominant structure of the dataset, while a limited number of companies exhibit significant deviations from the average profile. These deviations are primarily associated with extreme combinations of public procurement activity characteristics, such as a high number of active procedures, significant contract values, a large number of won contracts, or unusual values of metrics derived from public activity. These profiles align with the extreme companies previously identified by the PCA and clustering analyses, specifically Cluster 1 and Cluster 4.

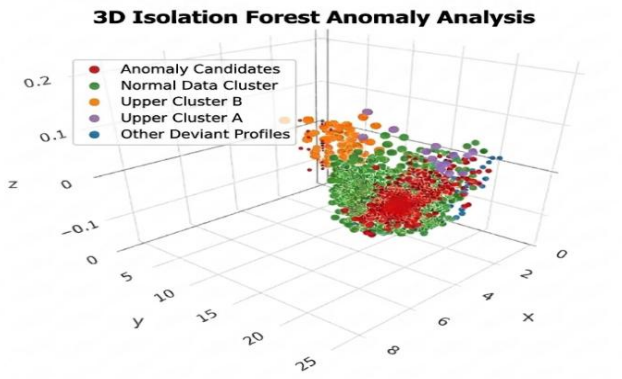

Fig. 3. Anomaly detection results using Isolation Forest.

Fig. 3 illustrates the distribution of companies based on the anomaly score generated by the Isolation Forest algorithm. It is observed that the majority of companies form a dense cluster near the average profile of the dataset, whereas isolated cases represent companies with characteristics that differ markedly from the rest of the analyzed population. These findings should not be interpreted as confirmed problematic cases, but rather as profiles requiring further investigation due to their deviation from the rest of the analyzed population.

Feature importance was analyzed using two tree-based ensemble models, Random Forest and Extra Trees [4], [30]. Although HistGradientBoosting achieved the highest performance in the classification benchmark, these two models were selected for the interpretability analysis due to their ability to estimate the relative contribution of input variables and the more direct interpretability of their decision tree structures. The feature-importance analysis indicated that variables related to public procurement activity contributed strongly to the models' discrimination between the heuristic classes. These importance measures are interpreted as indicators of variable contribution to predictive discrimination rather than as causal effects [12]. Among the most informative variables were the number of canceled procedures, the total value of the active budget limit, the number of active procedures, and other indicators describing the intensity and structure of the companies' public activity. This indicates that the operational information derived from procurement data contains the primary signal utilized by the models to differentiate company profiles.

Fig. 4 illustrates the relative importance of the ten most informative features according to the Random Forest model. These results should be interpreted as indicators of variable contribution to the classification process rather than as causal relationships between the features and the heuristic labels.

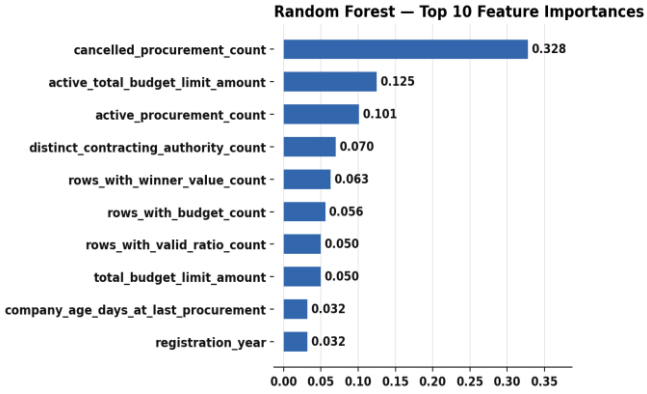

Fig. 4. Feature importance of the top predictive variables.

To assess the impact of financial information on model performance, an additional experiment was conducted on the subset of companies for which financial data from OpenCorporates was available. The objective of this experiment was to determine whether incorporating financial indicators provides supplementary information beyond the features extracted from public procurement activity.

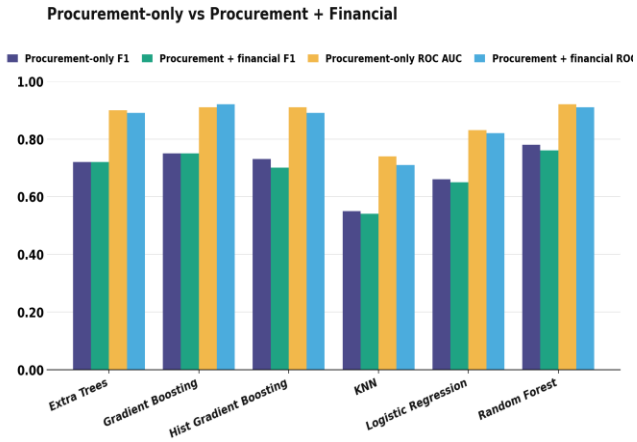

Fig. 5. Comparison of model performance with and without financial enrichment.

Unlike the feature importance analysis, which utilized only Random Forest and Extra Trees for interpretability, the financial enrichment experiment was conducted by comparing several of the primary classification models used during the benchmarking phase. This approach was employed to evaluate whether incorporating financial data yielded a consistent performance improvement across different models.

The results presented in Fig. 5 indicate that incorporating financial information did not lead to a significant improvement in classification performance. Although some models exhibited minor changes in evaluation metrics, the differences were insufficient to demonstrate a substantial increase in predictive power.

This finding suggests that, in the current dataset configuration, the information contained within public

procurement activity represents the primary source of signal for constructing the heuristic company profiles. The financial data serves as a complementary source, and its potential could be enhanced in future studies through the use of more comprehensive, up-to-date financial data with broader coverage.

V. Discussion

The results demonstrate that integrating APP procurement records with QKB administrative data at the company level provides analytical information that is not readily visible when procurement procedures are examined separately. The strong performance of the reduced-feature models (Table III) indicates that relevant discriminatory information remains after removing features directly used to construct the heuristic label. Similar applications have demonstrated the potential of machine learning for bidder classification and procurement monitoring [7], [33]. However, because the labels used in this study are heuristic, the results should be interpreted as company profiling rather than validated business-risk prediction.

The PCA and K-Means results (Table IV and Fig. 1 and 2) reveal substantial heterogeneity among economic operators. Most companies show relatively low procurement activity, while smaller groups exhibit substantially higher activity and contract exposure. From a procurement-governance perspective, these profiles can complement procedure-level monitoring by helping identify companies or patterns that may warrant further analytical review. Recent research and policy work similarly highlight the potential of machine learning and data analytics to support procurement oversight and risk-oriented monitoring [8], [35].

From a public-administration perspective, the integration of APP and QKB through the NIPT demonstrates the value of linking fragmented administrative information. The resulting company-level dataset provides a more comprehensive representation of economic operators and supports the analytical reuse of administrative data. This is consistent with recent developments in European public procurement, where interoperability, data integration, and reuse are increasingly emphasized [35], [37].

From a policy perspective, the anomaly-detection results (Fig. 3) suggest that integrated company-level indicators could support exploratory monitoring and prioritization of cases for further human review. Such use should not be understood as automated identification of misconduct or risk. Recent research emphasizes the value of combining analytical models with independently observed outcomes and complementary information sources for more reliable risk detection [35], [36].

The financial-enrichment experiment did not produce a substantial improvement in model performance. However, because financial information was available for only 3,159 companies (30.8% of the integrated dataset), this result should be interpreted cautiously. It indicates that procurement-derived variables provide the dominant analytical signal in the current dataset, rather than demonstrating that financial information has no value.

Overall, the findings support the analytical value of company-level integration of public procurement and administrative data in Albania. The main contribution lies in the empirical evidence enabled by integrating fragmented public information, while machine learning serves as an analytical tool for classification, company profiling, and anomaly detection rather than as the primary contribution.

VI. Conclusion

This study demonstrated the analytical value of integrating APP procurement records with QKB administrative business data at the company level through the NIPT identifier. The results showed that procurement-derived features contain substantial information for distinguishing heterogeneous company profiles, including atypical high-activity groups. The reduced-feature experiment and shuffled-label sanity check provided additional methodological support for the analysis.

The findings are relevant to procurement governance, public administration, and digital government, as integrated company-level data can support exploratory monitoring, company profiling, and the prioritization of cases requiring further institutional review. In the Albanian context, this company-level perspective complements existing research focused primarily on competition, savings, and accountability in public procurement [9].

However, the heuristic labels, predefined feature weights, limited financial-data coverage, and absence of independently validated risk outcomes restrict the interpretation of the models as formal business-risk predictors. Overall, the contribution lies in the empirical company-level analysis enabled by integrating fragmented public data in Albania, rather than in developing a new machine learning algorithm.

VII. Future Work

Future Research Should Focus on Independently Validated Risk Outcomes, Empirically Justified Feature Construction, Broader Longitudinal Financial Data, and Temporal and Graph-based Analysis of Relationships Among Companies and Contracting Authorities. Future Studies Could Also Examine the Interoperability of Additional Administrative Data Sources and Evaluate Whether Their Integration Improves Company-level Profiling. Explainable Artificial Intelligence (XAI) Methods Could Further Improve the Transparency and Interpretability of Model Outputs [34], [35], [36]

References

[1] J. Attard, F. Orlandi, S. Scerri, and S. Auer, “A systematic review of open government data initiatives,” Government Information Quarterly, vol. 32, no. 4, pp. 399–418, 2015.

[2] C. Batini and M. Scannapieco, Data and Information Quality: Dimensions, Principles and Techniques. Springer, 2016.

[3] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.

[4] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[5] T. Fawcett, “An introduction to ROC analysis,” Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[6] M. Lisciandra, R. Milani, and E. Millemaci, “A corruption risk indicator for public procurement,” European Journal of Political Economy, vol. 73, 102141, 2022. doi: 10.1016/j.ejpoleco.2021.102141.

[7] R. Nai, R. Meo, G. Morina, and P. Pasteris, “Public tenders, complaints, machine learning and recommender systems: A case study in public administration,” Computer Law & Security Review, vol. 51, 105887, 2023. doi: 10.1016/j.clsr.2023.105887.

[8] A. Salazar, J. F. Pérez, and J. Gallego, "VigIA: Prioritizing public procurement oversight with machine learning models and risk indices," Data & Policy, vol. 6, e75, 2024. doi: 10.1017/dap.2024.83.

[9] E. Martiri and E. Kapedani, "Competition, savings, and accountability: A study of Albania's public procurement of services," International Journal of Procurement Management, vol. 23, no. 1, pp. 106–125, 2025. doi: 10.1504/IJPM.2025.145556.

[10] M. Fazekas and G. Kocsis, "Uncovering high-level corruption: Cross-national objective corruption risk indicators using public procurement data," British Journal of Political Science, vol. 50, no. 1, pp. 155–164, 2020.

[11] M. Fazekas, I. J. Tóth, and L. P. King, "An objective corruption risk index using public procurement data," European Journal on Criminal Policy and Research, vol. 22, no. 3, pp. 369–397, 2016.

[12] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," Journal of Machine Learning Research, vol. 20, no. 177, pp. 1–81, 2019.

[13] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[14] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow, 3rd ed. O'Reilly Media, 2022.

[15] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, 2009.

[16] I. T. Jolliffe, Principal Component Analysis, 2nd ed. Springer, 2002.

[17] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1137–1143, 1995.

[18] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in Proceedings of the 8th IEEE International Conference on Data Mining, pp. 413–422, 2008.

[19] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297, 1967.

[20] R. Mitchell, Web Scraping with Python: Collecting More Data from the Modern Web, 2nd ed. O'Reilly Media, 2018.

[21] OECD, Preventing Corruption in Public Procurement. OECD Publishing, 2016.

[22] Open Contracting Partnership, Open Contracting Data Standard (OCDS), 2022.

[23] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," IEEE Data Engineering Bulletin, vol. 23, no. 4, pp. 3–13, 2000.

[24] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLOS ONE, vol. 10, no. 3, e0118432, 2015.

[25] World Bank, Enhancing Government Effectiveness and Transparency: The Fight Against Corruption. World Bank, 2020.

[26] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," Applied Artificial Intelligence, vol. 17, no. 5–6, pp. 375–381, 2003.

[27] Agjencia e Prokurimit Publik, "Export Public Calls," n.d. APP – Export Public Calls

[28] Qendra Kombëtare e Biznesit, "Kërko për subjekt," n.d. QKB – Kërko për subjekt

[29] OpenCorporates Albania, "Public Company Profile Pages by NIPT," n.d. OpenCorporates Albania

[30] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Machine Learning, vol. 63, no. 1, pp. 3–42, 2006.

[31] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987.

[32] T. M. Kodinariya and P. R. Makwana, "Review on determining number of clusters in K-means clustering," International Journal of Science and Engineering, vol. 1, no. 6, 2013.

[33] I. O. Silaschi, I. D. Pop, and A. M. Coroiu, "Improving transparency in Romanian public procurement: Machine learning to classify bidders and validate decisions," Procedia Computer Science, vol. 246, pp. 1210–1219, 2024. doi: 10.1016/j.procs.2024.09.547.

[34] T. Papadakis, I. T. Christou, C. Ipektsidis, J. Soldatos, and A. Amicone, "Explainable and transparent artificial intelligence for public policymaking," Data & Policy, vol. 6, e10, 2024. doi: 10.1017/dap.2024.3.

[35] OECD, *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions*. Paris, France: OECD Publishing, 2025, doi: 10.1787/795de142-en.

[36] M. Medina-Hernández, J. Kertész, and M. Fazekas, "Learning from sanctioned government suppliers: A machine learning and network science approach to detecting fraud and corruption in Mexico," *Scientific Reports*, vol. 16, Art. no. 22382, 2026, doi: 10.1038/s41598-026-48873-w

[37] European Commission, *Public Procurement Data Space (PPDS)*, European Commission, 2025–2026.