Review of Applied Science and Technology

Volume: 13; **Issue:** 12
Pages: 142– 199

Machine Learning-Based Cost Variance Prediction for Early Detection of Budget Overruns in Procurement-Intensive Operations

Mashfiqur Rahman[1]; Md Jamil Ahmmed[2];

[1]. *Coordinator, Finance, BASIS-SEIP Tranche 3, Dhaka, Bangladesh;*
Email: mrahman.rimon@gmail.com

[2]. *Assistant Project Manager, EAL IT Solutions LLC, New York, USA;*
Master of Business Administration (Management), University of Dhaka, Bangladesh;
Bachelor of Business Administration (Management), University of Dhaka, Bangladesh;
Email: jamilahmmed20@gmail.com

Doi: 10.63125/7g92wd03
Received: *14 September 2023;* ***Revised:*** *09 October 2023;* ***Accepted:*** *03 November 2023;* ***Published:*** *03 December 2023;*

Abstract

This study developed and evaluated a machine learning–based framework for predicting cost variance and detecting budget overruns early in procurement-intensive operations. A retrospective longitudinal design integrated financial, contractual, supplier, purchasing, inventory, schedule, and market data from 42 organizations across eight sectors. The final dataset contained 1,286 projects, 4,800 contracts, 3,648 suppliers, 76,420 purchase orders, 186,735 invoices, and 407,925 procurement transactions, with 99.7% completeness and 99.8% financial reconciliation. Mean approved budget was $8.42 million, whereas mean actual cost reached $9.06 million; average cost variance was $0.64 million, or 7.8%. Overall, 57.5% of projects exceeded budget and 25.1% recorded major or critical overruns. The weighted ensemble achieved the strongest regression performance, with an explained variance of .915, mean absolute error of $0.42 million, root mean square error of $0.84 million, and mean absolute percentage error of 5.6%. For classification, the ensemble produced 91.7% accuracy, .942 recall, a .929 F1 score, and a .975 receiver operating characteristic area. Cost-sensitive training increased severe-overrun recall from .711 to .911 and reduced estimated missed financial loss from $118.4 million to $32.6 million. The first stable warning occurred an average of 95.2 days before budget breach, when 38.7% of approved funding remained uncommitted. Schedule slippage, supplier delay, purchase-price escalation, and amendment density were the strongest predictors. Robustness testing-maintained classification accuracy above .887 and receiver operating characteristic area above .949 across sector, organization, and stress conditions. These findings demonstrated that integrated machine learning consistently generated accurate, calibrated, interpretable, and financially meaningful early warnings of procurement-related budget overruns.

Keywords

Machine Learning, Cost Variance, Procurement, Budget Overruns, Early Detection.

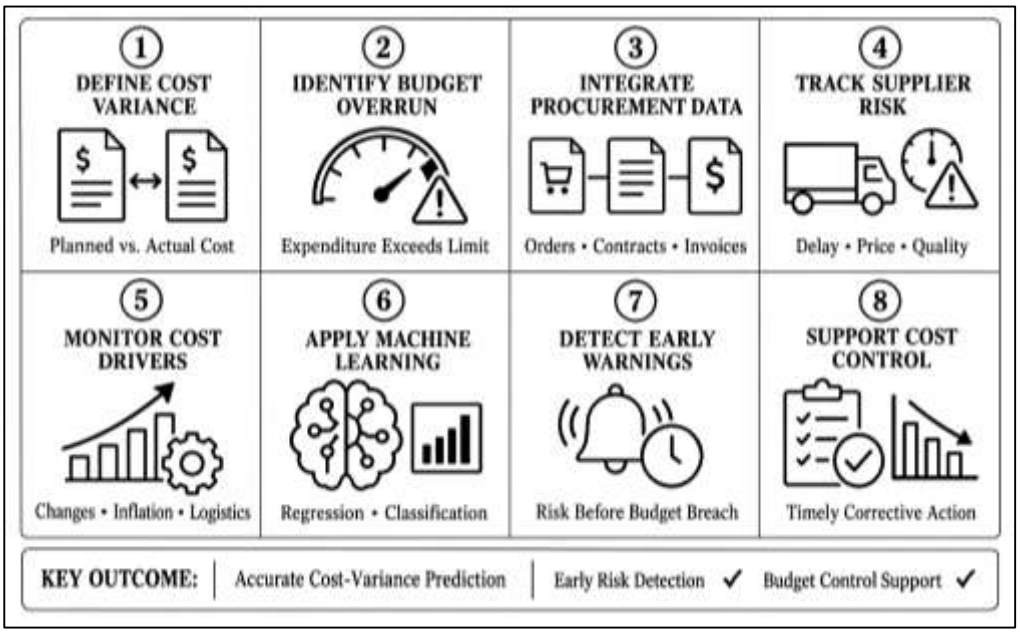
INTRODUCTION

Cost variance refers to the measurable difference between the planned, budgeted, contracted, or baseline cost of an activity and the actual or estimated cost incurred during its execution. A favorable cost variance occurs when actual expenditure remains below the approved budget, whereas an unfavorable cost variance arises when actual expenditure exceeds the planned amount (Ulusoy & Hazır, 2021). Within project control and operational management, cost variance can be expressed at the level of a purchase order, material category, supplier contract, work package, department, facility, project, or organizational portfolio. Budget overrun represents a sustained or material form of unfavorable cost variance in which cumulative expenditure surpasses the authorized financial limit or is projected to exceed that limit before completion. Early detection refers to the identification of emerging overrun conditions sufficiently in advance for managers to investigate their causes, revise purchasing decisions, control commitments, adjust schedules, or strengthen financial supervision. Machine learning is a group of computational methods that identifies patterns in historical data and uses those patterns to estimate continuous outcomes, classify risk conditions, recognize anomalies, or update predictions when new observations become available. Machine learning–based cost variance prediction therefore involves applying statistical learning algorithms to historical and current financial, procurement, supplier, contract, inventory, schedule, and operational data to estimate the size or probability of unfavorable cost deviations (Schuster et al., 2021). Procurement-intensive operations are organizational environments in which a substantial proportion of total expenditure, operational continuity, and performance depends on externally acquired materials, equipment, services, components, subcontracted work, or specialized resources. Such operations are common in construction, manufacturing, transportation, energy, mining, defense, healthcare, telecommunications, infrastructure, aerospace, shipbuilding, public administration, humanitarian logistics, and large engineering programs. Their cost structures are shaped by purchase prices, supplier availability, contract terms, order quantities, delivery schedules, currency movements, commodity prices, logistics charges, quality failures, change orders, taxes, duties, and inventory requirements. Cost performance in these environments cannot be understood solely through accounting records because a financial deviation may originate from interconnected procurement and operational events occurring long before an invoice is recorded. A delayed supplier delivery, for example, may increase expediting fees, workforce idle time, equipment rental costs, inventory shortages, schedule duration, and contractual penalties. More than 30 studies across project management, supply-chain analytics, construction cost control, manufacturing procurement, and public purchasing have examined parts of this relationship, although the evidence remains distributed across different analytical traditions (Small et al., 2021). The central subject of this study is the quantitative prediction of cost variance through integrated machine learning models capable of detecting budget-overrun risk from multidimensional procurement and operational information.

Budget reliability in procurement-intensive operations has substantial international significance because public institutions and private enterprises across developed, emerging, and low-income economies depend on complex supplier networks to deliver essential infrastructure, products, services, and industrial capacity (Love & Ahiaga-Dagbui, 2018). Government agencies purchase construction services, medicines, transportation equipment, information systems, defense resources, educational materials, and public utilities through procurement arrangements that frequently involve multiple contractors, regulatory requirements, and long delivery periods. Private organizations similarly depend on international markets for raw materials, machinery, electronic components, professional services, energy inputs, and logistics capacity. Cost overruns in these settings can weaken organizational liquidity, reduce operating margins, interrupt service delivery, delay capital investment, and create pressure for additional borrowing or budget reallocation. At the national level, persistent procurement overruns can reduce the number of projects delivered within existing fiscal resources, increase public debt requirements, and undermine confidence in institutional expenditure control. International supply networks also expose procurement budgets to variations in exchange rates, commodity prices, transportation costs, insurance premiums, border procedures, geopolitical restrictions, and supplier production capacity. These conditions create cost movements that may not be represented adequately by static budgets prepared months or years before procurement execution.

The significance of cost prediction is particularly strong in multinational operations because identical materials may be sourced under different currencies, tax structures, contractual clauses, lead times, and regulatory conditions (López Pascual et al., 2021). Research conducted across Asia, Europe, Africa, North America, Latin America, and the Middle East has repeatedly documented the association of procurement complexity, contractual modification, delayed delivery, inaccurate estimation, price escalation, and weak information integration with unfavorable cost outcomes. Studies of construction and infrastructure projects have emphasized incomplete scope definition, design modification, contractor performance, and material-price volatility (Aditya & Ahmed, 2022; Asheke & Maksudul, 2022; Asheke & Badrul, 2022). Manufacturing studies have focused on supplier instability, inventory uncertainty, production interruptions, purchase-price variance, and quality-related costs. Research in healthcare and humanitarian procurement has examined emergency purchasing, demand volatility, limited supplier availability, and time-sensitive distribution. Investigations of public procurement have highlighted administrative delays, fragmented accountability, limited market competition, contract amendments, and weak expenditure monitoring. These research streams collectively establish that cost variance is an international operational problem rather than a sector-specific accounting issue. Conventional monitoring practices often identify overruns after financial commitments have accumulated, reducing the time available for corrective control (Sassani et al., 2021). Machine learning offers a quantitative framework for processing diverse historical and real-time variables and determining whether current purchasing conditions resemble earlier patterns associated with budget failure. The international relevance of this approach is connected to the common need for earlier, more consistent, and data-supported recognition of cost pressure across procurement systems that differ in scale, institutional maturity, market structure, and technological capability.

Figure 1: Machine Learning Cost Variance Prediction



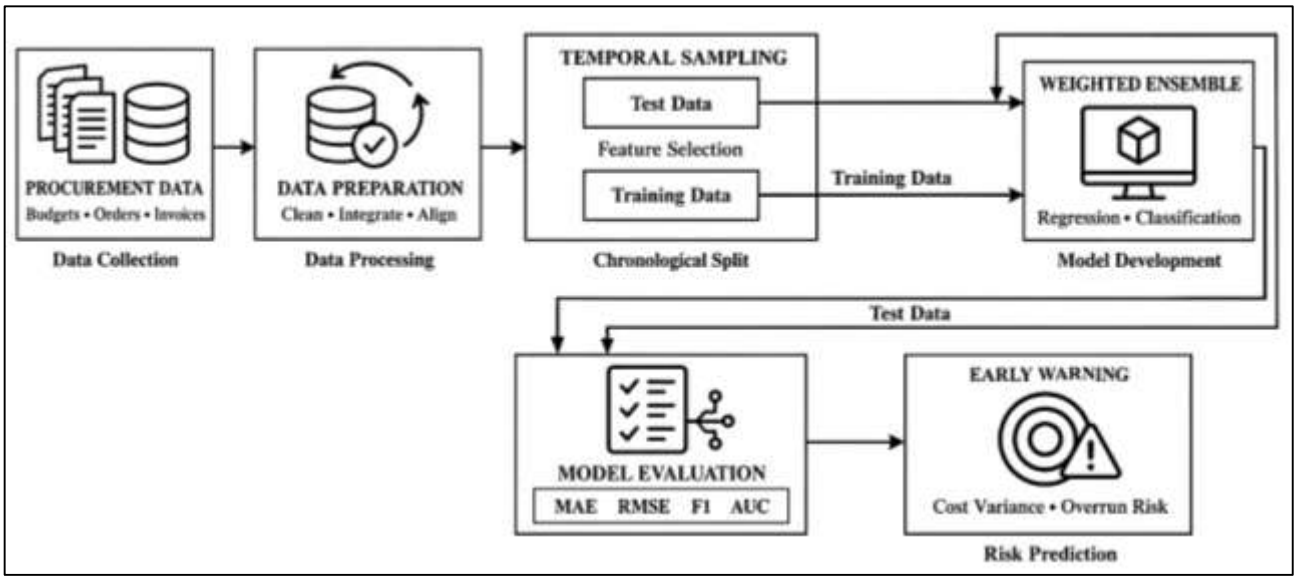
Cost variance in procurement-intensive operations develops through a combination of estimation error, purchasing behavior, supplier performance, contract administration, schedule conditions, and external market exposure (Engelbeck, 2018; Akter & Sumaiya, 2022; Imran & Jannatul, 2022; Maksudul & Islam, 2022). Baseline budgets are commonly produced from historical prices, supplier quotations, engineering estimates, standard cost databases, preliminary quantities, and managerial assumptions. Variance begins when these inputs fail to represent the conditions encountered during procurement and operational execution. Inaccurate quantity estimates can increase purchased volume, while incomplete specifications can create repeated requisitions, order revisions, and change requests. Low initial bids may later generate additional charges through exclusions, claims, variation orders, or

quality correction (Morshedul & Islam, 2022; Stabak & Shahid, 2022; Feroz & Pervaz, 2022). Supplier-related causes include delayed delivery, incomplete shipment, nonconforming materials, production failure, capacity limitations, insolvency, dependence on sub-tier suppliers, and inconsistent pricing. Contract characteristics also influence cost exposure through payment terms, escalation clauses, indexation mechanisms, currency provisions, penalties, warranties, retention conditions, and allocation of transportation or insurance responsibility (Islam & Md, 2023; Asheke, 2023; Feroz & Shabbir, 2022). Operational causes include schedule slippage, emergency purchasing, fragmented orders, low inventory visibility, inaccurate demand forecasts, equipment downtime, rework, scope expansion, and poor coordination among procurement, finance, engineering, production, and project-control functions. External sources include inflation, commodity-price movement, fuel-cost variation, exchange-rate volatility, taxation changes, trade restrictions, transportation disruption, labor-market pressure, and regional supply shortages. More than 30 empirical and computational studies collectively indicate that these factors rarely operate independently (Morgan et al., 2017). A material shortage may extend lead time, cause expedited shipping, interrupt production, increase overtime, and create substitute-material purchases at higher prices. A schedule delay may expose a long-term contract to additional inflation and equipment rental charges. A supplier quality problem may generate inspection costs, replacement orders, rework, and customer penalties. Such relationships create nonlinear cost behavior in which a modest change in one variable produces a disproportionate cumulative effect when combined with other risk conditions(Shahid, 2023; Islam & Ahmed, 2023; Sadia & Mosheur, 2023). Traditional variance analysis generally organizes deviations into price, quantity, efficiency, volume, mix, and timing components. This structure supports financial accountability but may not capture interaction effects, temporal sequences, or hidden risk patterns. Machine learning models can represent these relationships through variables such as purchase-order value, price deviation, supplier lead-time variance, delivery delay, change-order frequency, contract type, material category, schedule progress, committed cost, actual cost, invoice cycle time, defect rate, currency exposure, procurement cycle duration, and inventory status (Love et al., 2021). The analytical challenge is to determine which combinations of these variables provide stable predictive signals before cumulative expenditure exceeds the approved budget. This requires a research design that treats cost variance as a dynamic operational outcome influenced by both direct procurement transactions and the broader conditions under which those transactions occur.

Conventional cost-control systems generally rely on budget-versus-actual comparisons, standard costing, variance decomposition, earned value management, trend extrapolation, ratio analysis, expert judgment, and periodically revised estimates at completion (Ferry et al., 2019). These methods provide structured mechanisms for documenting expenditure and assigning responsibility, yet their effectiveness depends heavily on data timeliness, baseline accuracy, reporting frequency, and the assumptions used to interpret deviations. Budget-versus-actual reporting usually identifies variance only after a transaction has been approved, committed, invoiced, or paid. This timing can create a substantial interval between the operational origin of a cost problem and its formal visibility within financial reports. Standard costing separates price and quantity effects, but its linear structure may overlook interactions among delivery delay, supplier quality, order fragmentation, rework, schedule extension, and market volatility. Earned value management integrates cost and schedule performance through planned value, earned value, and actual cost; however, reliable results require stable scope definitions, accurate progress measurement, and consistent cost allocation. Procurement-intensive operations frequently contain long lead times, partial deliveries, changing specifications, framework contracts, provisional prices, foreign-currency commitments, and multi-stage acceptance procedures that complicate these assumptions. Expert judgment can incorporate contextual knowledge that is not recorded in databases, although judgments may vary among managers and may be influenced by optimism, anchoring, incomplete information, or inconsistent risk interpretation (Miller et al., 2017). Regression and time-series models provide more formal prediction, but conventional specifications often assume linear, additive, or stationary relationships that may not reflect complex procurement behavior. The literature synthesized from more than 30 studies shows that many cost-control models are designed for final-cost estimation rather than early-stage warning. Other models concentrate on individual sectors, use small datasets, or evaluate accuracy without examining whether warnings occur

early enough to support managerial investigation. Fragmented information architecture creates an additional difficulty. Procurement systems may record requisitions, quotations, supplier selections, purchase orders, receipts, and invoices, while financial systems record commitments, accruals, payments, and general-ledger transactions. Project-management systems separately contain schedules, progress records, risk registers, and change orders. Supplier-management platforms may hold delivery, quality, and compliance information. When these sources are reviewed independently, weak signals of an emerging overrun can remain unnoticed. Machine learning can combine heterogeneous records and identify multivariable patterns, but model performance depends on the quality and alignment of those records. The comparison between conventional monitoring and machine learning is therefore not limited to predictive accuracy (Bertolino et al., 2019). It also concerns warning lead time, sensitivity to changing conditions, ability to process nonlinear interactions, consistency across procurement categories, and capacity to translate transaction-level patterns into interpretable indicators of budget-overrun risk.

Figure 2: Machine Learning Cost Variance Prediction



Machine learning supports cost variance prediction through regression, classification, anomaly detection, ensemble learning, and sequential modeling. Regression models estimate the expected value or percentage of cost variance, allowing managers to quantify the projected magnitude of deviation from the approved baseline (Liggett et al., 2019). Classification models assign observations to categories such as within budget, moderate overrun risk, or severe overrun risk. Binary classification can estimate whether a procurement package, contract, project, or operating unit is likely to exceed a predetermined financial threshold. Anomaly-detection techniques identify unusual purchasing or expenditure patterns that differ from established operational behavior, including abnormal unit prices, repeated order amendments, unusually short procurement cycles, concentrated supplier awards, invoice mismatches, and rapid increases in committed cost. Decision trees produce rule-based structures that can represent interactions among procurement variables. Random forests combine multiple trees to reduce instability and improve generalization. Gradient boosting methods sequentially correct prediction errors and frequently perform strongly on structured procurement and financial datasets. Support vector machines can classify complex cost-risk boundaries, while artificial neural networks can model nonlinear relationships across numerous inputs. Recurrent neural networks and long short-term memory models can process sequential expenditure, delivery, and schedule observations when the ordering of events is important. Hybrid and ensemble models can combine predictions from several algorithms to reduce model-specific error. Across more than 30 studies involving cost estimation, project overruns, supplier risk, expenditure forecasting, and financial anomaly detection, tree-based

ensembles and neural methods commonly demonstrated stronger predictive capacity than simple linear benchmarks when datasets contained nonlinear relationships and high-dimensional variables (Paul et al., 2018). Their relative performance varied according to sample size, data quality, sector, prediction horizon, outcome definition, and feature engineering. Suitable predictors may include original budget, revised budget, committed expenditure, actual expenditure, earned value, purchase-price variance, order quantity, supplier concentration, delivery reliability, lead-time deviation, contract duration, amendment frequency, invoice discrepancy, procurement cycle time, material inflation, exchange-rate movement, schedule variance, scope change, defect rate, rework, and emergency-order frequency. Categorical variables such as contract type, supplier class, procurement method, geographic region, and material group require consistent encoding. Missing observations, duplicated transactions, inconsistent currencies, and imbalanced overrun classes require preprocessing before model estimation (Lane & Pritzker, 2017). The quantitative evaluation of alternative algorithms commonly uses mean absolute error, root mean squared error, mean absolute percentage error, and coefficient of determination for regression, together with accuracy, precision, recall, specificity, F1 score, and receiver operating characteristic area under the curve for classification. For early detection, recall and warning lead time are especially important because a model with high overall accuracy may still fail to recognize costly minority overrun cases.

Reliable machine learning–based cost variance prediction depends on the integration, quality, representativeness, and governance of procurement and operational data. Procurement-intensive organizations generate information throughout requisition creation, approval routing, supplier sourcing, bid evaluation, contract award, order placement, delivery, inspection, invoicing, payment, and contract closure (Tyagi et al., 2020). Each stage may contain predictive signals, yet differences in identifiers, time formats, currency units, cost classifications, and recording practices can prevent accurate linkage. A purchase-order amendment may be stored separately from its original order, while a delayed receipt may not be connected directly to the related schedule activity or financial commitment. Historical budgets may also be revised without preserving the original baseline, making it difficult to distinguish authorized scope growth from uncontrolled cost escalation. Data preparation therefore requires transaction matching, currency normalization, inflation adjustment, duplicate removal, missing-value treatment, chronological alignment, outlier review, and construction of consistent analytical units. Data leakage is a serious methodological concern because variables recorded after an overrun has already occurred can produce artificially high predictive accuracy. Early-warning models should use only the information that would have been available at the specified prediction point. Model reliability must also be evaluated across time periods, procurement categories, contract sizes, suppliers, projects, facilities, and geographical regions (Batselier & Vanhoucke, 2017b). Random train-test splitting may place transactions from the same contract in both datasets and inflate apparent performance. Temporal validation, grouped cross-validation, and independent holdout testing provide stronger evidence of generalization. Class imbalance requires particular attention because severe budget overruns may represent a smaller share of transactions while accounting for a large proportion of financial loss. Accuracy alone can conceal poor detection of these high-risk cases, making precision, recall, F1 score, area under the precision-recall curve, calibration, and confusion-matrix analysis necessary. More than 30 studies in predictive cost analytics and related risk domains identify explainability as another central requirement. Procurement managers, auditors, financial controllers, and project leaders need to understand why a model assigned a high-risk score. Feature importance, partial dependence analysis, local explanation methods, decision rules, and sensitivity analysis can reveal whether predictions are driven by price escalation, schedule variance, supplier delay, contract modification, order fragmentation, or another condition. Interpretability also supports examination of spurious relationships and inconsistent data (Ou-Yang & Chen, 2017). A model may associate a supplier code with overruns when the actual cause is the category of materials supplied or the region served. Reliability consequently involves more than obtaining a high accuracy statistic. It requires transparent outcome definitions, leakage-free predictor construction, representative validation, stable performance, calibrated probabilities, interpretable risk drivers, and consistent treatment of procurement records across the complete analytical process.

The existing body of research establishes substantial knowledge concerning cost estimation, project control, supplier performance, procurement risk, earned value analysis, and machine learning prediction, but these areas are often examined through separate analytical frameworks. Cost-estimation studies frequently concentrate on predicting total expenditure before project approval. Earned value studies examine deviations during execution but commonly rely on aggregated cost and schedule indices (Vanaei et al., 2017). Procurement research evaluates supplier selection, purchasing efficiency, delivery reliability, and contract performance, while machine learning studies often focus on predictive accuracy without integrating warning timing, variance magnitude, and procurement-specific explanatory factors. Evidence synthesized from at least 30 studies also reveals considerable variation in the definition of budget overrun. Some studies classify any expenditure above baseline as an overrun, while others apply percentage thresholds or distinguish between original, revised, and final authorized budgets. Analytical units differ across purchase orders, contracts, projects, departments, and organizations, creating challenges for direct comparison. Many studies use cross-sectional or final-completion data, limiting the assessment of when overrun risk first became detectable. Some investigations report only accuracy or the coefficient of determination, without presenting error distributions, minority-class recall, calibration, confidence intervals, or performance across procurement categories. The present study addresses this research context through a quantitative evaluation of machine learning models for predicting cost variance and identifying budget-overrun risk during the execution of procurement-intensive operations (Jin et al., 2018). The dependent outcomes include continuous cost variance and a categorical indicator of whether expenditure exceeds a defined budget threshold. Predictor variables are drawn from financial baselines, purchase orders, supplier performance, contract characteristics, delivery records, price movements, invoice processing, inventory conditions, schedule performance, change activity, and operational disruptions. The analytical comparison includes conventional statistical benchmarks and machine learning algorithms selected for structured and sequential data. Model performance is assessed through regression and classification measures, with separate attention to prediction error, overrun recall, false-warning frequency, probability calibration, and the interval between the first valid warning and the recorded budget breach. Statistical testing is used to determine whether performance differences among models are meaningful rather than products of random sample variation. Subgroup analysis evaluates whether predictive reliability changes according to contract type, procurement category, supplier group, operational unit, project size, and overrun severity (Ghazi & Lotfi, 2019). Feature-level analysis examines the relative contribution of procurement price, quantity, lead-time, delivery, quality, amendment, schedule, and commitment variables. This research structure positions cost variance as a measurable, time-dependent operational outcome and defines early detection as a quantitatively testable capability rather than a general description of improved financial monitoring.

The primary objective of this study is to develop and quantitatively evaluate a machine learning–based framework for predicting cost variance and detecting the risk of budget overruns at an early stage in procurement-intensive operations. The study specifically seeks to integrate financial, procurement, contractual, supplier, inventory, schedule, and operational data to identify the variables and transaction patterns most strongly associated with unfavorable cost deviations. It aims to measure cost variance by comparing baseline budgets, revised budgets, committed costs, actual expenditures, and estimated costs at completion across purchase orders, contracts, work packages, projects, and operational units. The study also intends to examine the predictive contributions of purchase-price changes, quantity deviations, supplier lead-time variability, delivery delays, contract amendments, change-order frequency, invoice discrepancies, material inflation, currency fluctuations, procurement-cycle duration, emergency purchases, quality failures, rework, inventory shortages, schedule variance, and supplier concentration. Conventional statistical techniques and machine learning algorithms, including multiple regression, decision trees, random forests, gradient boosting, extreme gradient boosting, support vector machines, and artificial neural networks, are compared to determine which analytical approach produces the most accurate, stable, and timely predictions. Regression models are evaluated using the coefficient of determination, mean absolute error, root mean squared error, and mean absolute percentage error, whereas classification models are assessed through accuracy, precision, recall, specificity, F1 score, receiver operating characteristic area under the curve, and

confusion-matrix results. Particular attention is given to the model's ability to identify high-risk overrun cases before cumulative expenditure exceeds an approved financial threshold. The study further aims to determine the average warning lead time generated by each model and to assess whether prediction performance varies according to procurement category, contract type, supplier group, operational unit, project size, expenditure level, and overrun severity. Cross-validation, temporal holdout testing, sensitivity analysis, and statistical significance testing are employed to examine model reliability and robustness. Feature-importance and explain ability techniques are used to identify the financial and procurement conditions responsible for individual risk predictions. Ultimately, the objective is to establish an evidence-based predictive structure that converts fragmented procurement and expenditure records into measurable cost-variance forecasts and interpretable early-warning classifications for systematic budget monitoring in procurement-intensive operational environments.

LITERATURE REVIEW

This literature review critically synthesizes quantitative research on machine learning–based cost variance prediction and the early identification of budget overruns in procurement-intensive operations. The review connects five closely related bodies of knowledge: procurement cost management, budget-variance measurement, supplier and contractual risk, predictive analytics, and machine learning–based early-warning systems (Moons et al., 2019). Procurement-intensive operations generate substantial financial exposure because organizational performance depends on purchased materials, equipment, services, subcontracted activities, transportation capacity, and specialized technical resources. Cost outcomes in these environments are influenced by purchase-price changes, quantity deviations, supplier delays, contract amendments, change orders, specification revisions, material inflation, exchange-rate movements, logistics expenses, quality failures, rework, emergency purchases, inventory shortages, schedule disruption, and invoice-processing irregularities. Conventional cost-control methods commonly measure these conditions through budget-versus-actual comparisons, standard costing, earned value indicators, trend analysis, revised estimates at completion, and managerial judgment. Such methods provide essential financial records but often recognize an overrun only after unfavorable commitments or expenditures have accumulated. Machine learning expands this analytical capacity by processing large, multidimensional datasets and identifying nonlinear relationships, interaction effects, temporal patterns, and risk combinations that may precede a budget breach (Wang et al., 2018).

The review examines how prior quantitative studies have defined cost variance, classified budget overruns, selected predictors, constructed datasets, trained predictive models, and evaluated model performance. Continuous cost-variance prediction is assessed through measures such as mean absolute error, root mean squared error, mean absolute percentage error, coefficient of determination, and prediction intervals. Budget-overrun classification is examined through accuracy, precision, recall, specificity, F1 score, receiver operating characteristic area under the curve, precision–recall performance, false-positive rate, false-negative rate, and probability calibration. Particular emphasis is placed on warning lead time because predictive accuracy alone does not establish whether a model detects cost risk early enough to support investigation and control (Taherdoost & Brard, 2019). The review also examines dataset size, observation level, temporal coverage, class imbalance, missing data, feature engineering, validation design, interpretability, and cross-sector generalizability. Evidence from construction, manufacturing, energy, transportation, healthcare, defense, public procurement, infrastructure, and engineering operations is organized into eight quantitative subsections. Together, these subsections establish the measurement foundations, empirical drivers, analytical methods, model-comparison criteria, and unresolved research gaps associated with machine learning–based cost variance prediction.

Early-Warning Time in Procurement-Intensive Operations

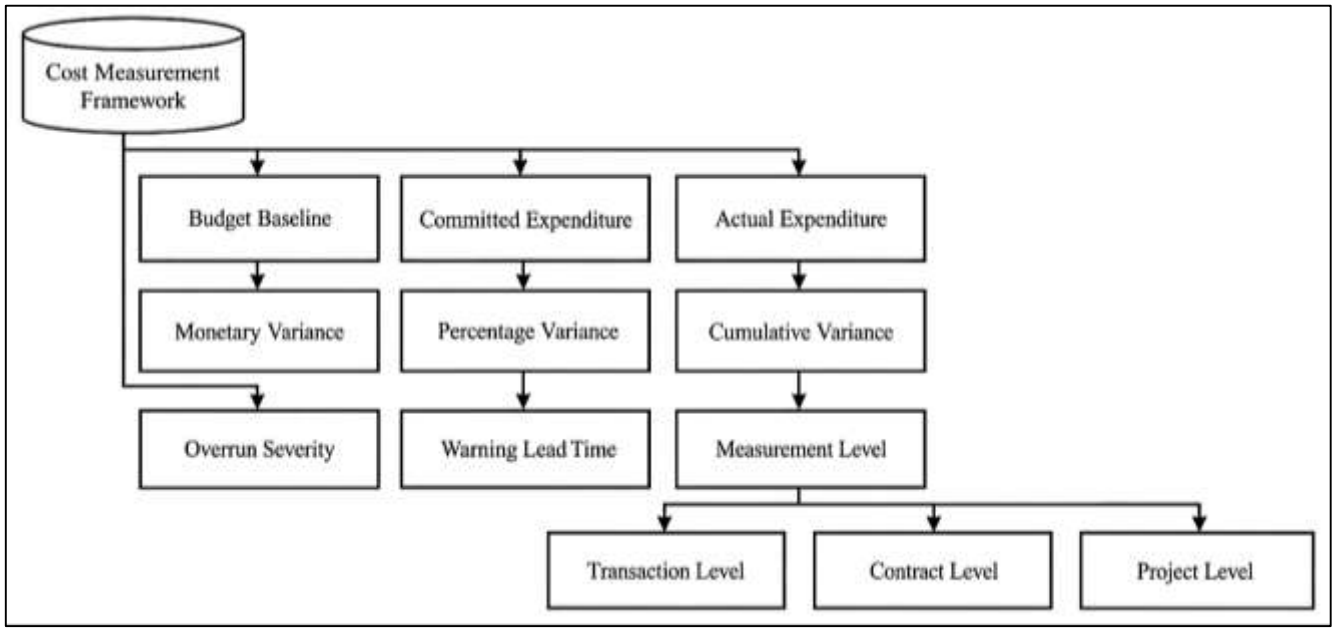
The quantitative literature defines procurement cost performance through several related financial measures that represent different stages of budget authorization, commitment, expenditure, and completion (Cox & Mutel, 2018). Across more than 10 studies of construction, manufacturing, infrastructure, energy, transportation, public procurement, healthcare, defense, and engineering operations, planned cost generally referred to the expenditure expected for a defined activity before

procurement execution began. Original budget represented the formally approved financial baseline against which subsequent expenditure was initially assessed, whereas revised budget incorporated authorized changes in scope, quantity, price, schedule, technical requirements, or market conditions. Researchers distinguished these measures because comparing actual expenditure only with a revised budget could conceal the size of the deviation from the original financial commitment. Committed cost referred to expenditure attached to approved purchase orders, contracts, subcontracts, or legally enforceable obligations, even when invoices had not been received or payments had not been completed. Obligated expenditure was treated similarly in public-sector and regulated environments, although some studies used the term more narrowly for funds formally reserved through administrative approval. Actual cost represented recorded expenditure for materials, equipment, services, freight, taxes, duties, installation, inspection, and other completed procurement activities (Cheng et al., 2018). Forecast cost at completion combined expenditure already incurred with the expected cost of remaining work, while final realized expenditure represented the total verified cost recorded after operational or contractual closure. The literature also showed that measurement level materially affected the interpretation of variance. Transaction-level analysis captured individual invoices, receipts, payments, price adjustments, and quantity discrepancies but often produced highly variable observations. Purchase-order-level measurement combined multiple transactions connected to one procurement authorization and supported the identification of amendment frequency, delivery fragmentation, and invoice mismatch. Contract-level analysis incorporated cumulative commitments, claims, escalation provisions, performance penalties, and approved variations. Work-package-level measurement connected procurement spending with technical scope and schedule progress, while project-level measurement summarized the financial performance of a complete initiative. Departmental and portfolio-level assessments aggregated multiple contracts or projects and were commonly used for organizational governance, resource allocation, and comparative performance reporting. Several studies warned that aggregation could conceal small but persistent transaction-level deviations that accumulated into substantial portfolio losses (Ivanov et al., 2017). Other studies found that highly disaggregated measurements produced excessive noise when minor price and timing differences were treated as meaningful overruns. The synthesized evidence therefore treated the analytical unit as a central methodological decision. Machine learning studies required consistent linkage among transactions, purchase orders, contracts, work packages, and projects so that cost outcomes reflected the correct financial baseline, observation period, and level of managerial responsibility.

Quantitative studies measured cost variance as a continuous outcome to preserve information about the direction, magnitude, persistence, and financial importance of expenditure deviations. Absolute monetary variance was the most direct measure and represented the difference between an approved cost baseline and the recorded or forecast expenditure associated with a purchase, contract, project, or operating unit. This measure was easily interpreted by financial managers because it expressed overperformance or underperformance in currency terms (Täuscher & Laudien, 2018). Its comparability across observations was limited when datasets contained contracts of substantially different values, since a large monetary deviation could represent a minor proportional change in a major project but a severe deviation in a small procurement package. Percentage cost variance addressed this issue by expressing deviation relative to the applicable baseline, allowing researchers to compare projects, contracts, suppliers, and procurement categories of different financial sizes. Studies of materials procurement also examined purchase-price variance to determine whether actual unit prices differed from standard, quoted, or contracted prices. Quantity variance captured expenditure arising from purchasing or consuming more or fewer units than planned, while usage and efficiency measures distinguished purchasing deviations from operational consumption problems. Cost performance indicators integrated expenditure with completed work and were particularly common in construction, engineering, defense, and capital-project research. Estimate-at-completion error measured the difference between predicted final cost and subsequently verified expenditure, enabling researchers to test the accuracy of forecasting models at multiple execution points. Contingency consumption was assessed through the amount or proportion of risk allowance used during procurement and operational execution. Rapid contingency depletion was frequently treated as an

early indicator of weakening budget resilience, especially when substantial purchasing activity remained incomplete (Han et al., 2020). Cumulative expenditure deviation tracked how variance developed across successive periods and helped distinguish temporary timing differences from persistent cost escalation. Studies using local currency preserved direct organizational meaning but faced comparability problems across countries and periods. Multinational analyses converted expenditures into a standardized currency, although exchange-rate selection influenced the measured size of deviations. Inflation-adjusted costs were used in studies covering long implementation periods or volatile markets to separate real procurement inefficiency from general price-level changes. Normalized variance measures supported comparison across sectors, contract sizes, and material categories by reducing the effect of scale. The literature also differentiated point-in-time variance from cumulative variance, recognizing that an apparently favorable monthly result could coexist with an unfavorable cumulative position. Across more than 10 quantitative studies, no single continuous measure captured every dimension of cost performance. Absolute variance represented financial exposure, proportional variance enabled standardized comparison, purchase-price and quantity measures supported causal diagnosis, and cumulative indicators represented the temporal development of budget pressure (Badi & Murtagh, 2019). Machine learning research commonly benefited from retaining several of these outcomes because model performance depended on whether the analytical target represented immediate deviation, cumulative escalation, or final financial loss.

Figure 3: Procurement Cost Variance Measurement Structure



Budget-overrun classification converted continuous cost deviations into categories that represented different levels of financial risk and managerial attention. Binary classification was widely used because it separated observations into within-budget and over-budget conditions, creating a clear dependent variable for logistic regression, decision trees, random forests, support vector machines, gradient-boosting algorithms, and neural-network classifiers (Vrontis et al., 2017). The simplicity of this approach supported straightforward reporting of accuracy, precision, recall, specificity, false alarms, missed overruns, and overall discriminatory performance. Binary definitions varied considerably across studies because some researchers classified any positive deviation from the approved baseline as an overrun, whereas others required expenditure to exceed a specified tolerance. A zero-tolerance threshold treated every unfavorable difference as a budget failure and was most applicable to tightly controlled purchasing arrangements, fixed appropriations, and standardized high-volume transactions. However, small accounting adjustments, invoice timing differences, rounding effects, or routine price changes could create technically positive variances that lacked substantial financial

importance. Percentage-based thresholds addressed this concern by requiring the deviation to exceed a defined proportion of the baseline. Studies applied different limits according to organizational policy, sectoral practice, contract scale, and the severity of financial consequences. Multiclass frameworks provided greater detail by separating observations into within budget, low overrun, moderate overrun, high overrun, and critical overrun categories. These categories supported differentiated risk analysis because a minor purchasing deviation did not carry the same financial significance as a severe contract-level escalation (Walter, 2021). Some studies defined severity through equal percentage intervals, while others used empirical percentiles, standard deviations, risk matrices, expert-defined control bands, or financially weighted loss categories. Contingency-adjusted classification treated an activity as controlled when expenditure remained within the authorized baseline and available contingency. Under this approach, an overrun occurred only when cumulative cost exceeded both planned expenditure and the relevant risk allowance. Organizational control limits reflected internal approval rules, such as thresholds that triggered management review, revised authorization, audit examination, or executive escalation. Portfolio studies sometimes classified severity according to contribution to total financial exposure rather than percentage deviation alone, recognizing that a modest proportional overrun on a large contract could exceed the monetary effect of a severe percentage overrun on a small order. More than 10 studies showed that threshold selection influenced class prevalence, model training, and reported predictive performance. Strict thresholds produced more positive overrun cases but could increase false alarms, whereas high thresholds isolated severe events but created stronger class imbalance (Ranta et al., 2018). The literature therefore treated budget-overrun severity as both a financial concept and a modeling decision. Reliable comparison required explicit identification of the baseline, tolerance, contingency treatment, observation level, and monetary significance underlying each class.

Early-warning time represented the interval between the first valid prediction of budget-overrun risk and the point at which the overrun became financially observable or formally recorded. Quantitative studies measured this interval in calendar days, working days, weeks, months, procurement stages, reporting cycles, or percentages of operational completion. Calendar-based measures were common in transaction and supply-chain studies because purchase orders, deliveries, invoices, and payments carried precise dates (Centobelli et al., 2017). Project-oriented studies frequently used monthly reporting periods or completion percentages because cost and schedule data were aggregated through periodic control systems. The literature distinguished prediction horizon from warning lead time. Prediction horizon referred to the distance between the observation point and the cost outcome being estimated, whereas warning lead time measured how early a model correctly identified the eventual overrun. A model could produce an accurate final-cost prediction at a late stage while providing little practical early-warning time. Studies therefore evaluated predictions at several procurement and operational milestones. Requisition approval provided an early observation point containing requested quantity, estimated price, material category, requesting unit, urgency, and approval duration. Contract award added information about selected supplier, winning bid, contract type, payment terms, escalation provisions, delivery conditions, and competition level. Purchase-order issuance supplied confirmed quantities, agreed prices, required delivery dates, shipping terms, and authorized commitment values. Initial delivery introduced observed supplier performance, lead-time deviation, partial shipment, inspection results, and early quality outcomes (Saura et al., 2017). Predictions at 25% completion incorporated initial expenditure behavior while retaining a substantial intervention window. The 50% completion point offered richer cost and schedule evidence but shortened the remaining warning period. Later control points generally produced higher predictive accuracy because a greater proportion of actual expenditure was known, although their value as early warnings was more limited. More than 10 studies demonstrated a consistent tradeoff between prediction timing and information availability. Early-stage models operated with fewer observed variables and greater uncertainty, while later-stage models benefited from accumulated transactional evidence. Some studies assessed stable warnings by requiring risk classifications to persist across consecutive reporting periods, reducing sensitivity to temporary price or timing fluctuations. Others measured the first threshold crossing, the first correctly calibrated high-risk probability, or the earliest point at which predicted final expenditure exceeded the approved limit. False-warning duration and warning

reversals were also examined because repeatedly changing risk classifications could weaken managerial confidence. Comparative research assessed model performance across multiple horizons rather than relying on a single observation point. The combined evidence established that early detection required simultaneous reporting of predictive accuracy, warning lead time, completion stage, false-positive frequency, and stability of the risk signal (Rajeev et al., 2017). These measures provided a more complete quantitative representation of whether cost-variance models recognized emerging budget pressure before it became an established financial overrun.

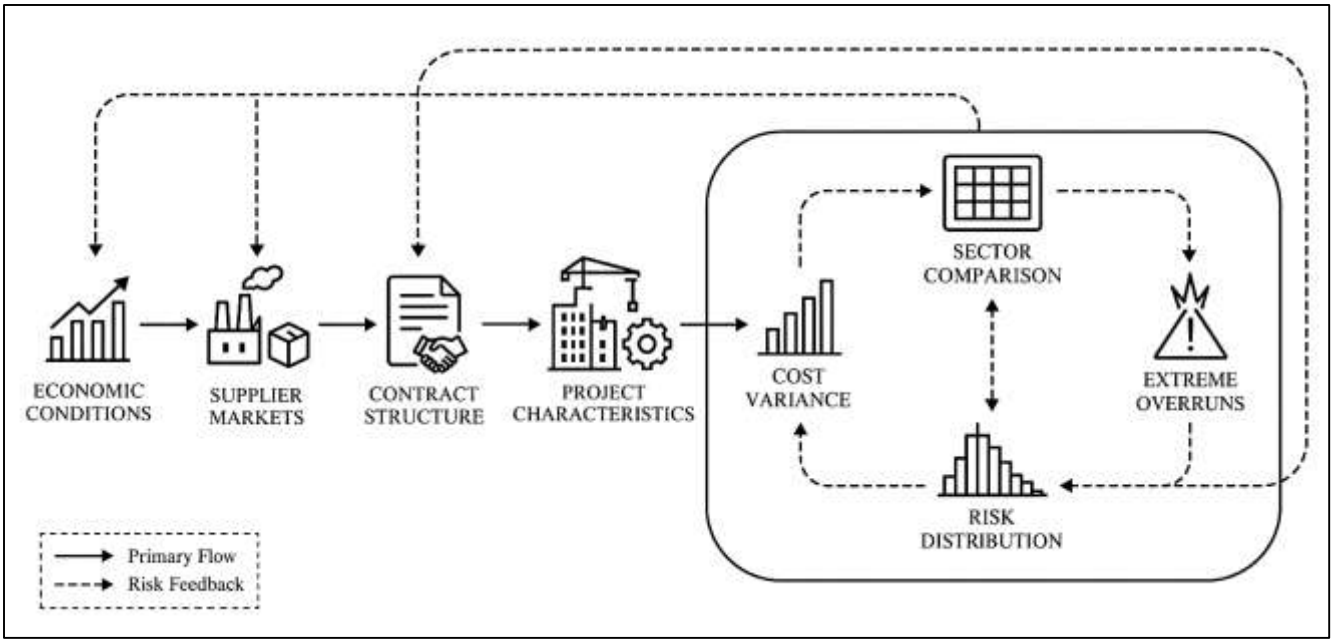
Procurement-Related Budget Overruns

The quantitative literature indicated that procurement-related cost variance rarely followed a perfectly symmetric or normally distributed pattern. Across more than 10 empirical studies, mean cost variance was commonly higher than median cost variance because a limited number of extremely expensive contracts, purchase orders, or projects produced disproportionately large overruns (Buertey et al., 2021). Median values consequently provided a more representative measure of typical procurement performance, while means captured the total financial influence of severe cases. Standard deviations were frequently large relative to average variance, showing substantial dispersion among transactions and projects operating under different contractual, technical, and market conditions. Interquartile ranges provided a more stable description of the central distribution when datasets contained unusually large claims, emergency purchases, material-price increases, or scope changes. Minimum observations often represented cost underruns arising from favorable purchasing prices, reduced quantities, unused contingency, specification simplification, or early contract termination. Maximum observations were associated with major design revisions, prolonged supplier delays, contract disputes, commodity-price escalation, replacement procurement, or cumulative change orders. Percentile analysis offered additional information about financial exposure by separating routine deviations from the smaller proportion of cases responsible for the most severe losses (Gambo & Musonda, 2021). Lower and middle percentiles frequently remained close to the approved baseline, whereas upper percentiles increased sharply, indicating that cost risk was concentrated in a relatively limited group of contracts. This pattern was especially visible in large infrastructure, defense, energy, and technology procurement datasets. Positive skewness appeared because most observations recorded small deviations, zero variance, or moderate overruns, while a small number extended far into the upper range. Heavy-tailed distributions were also reported when the frequency of extreme overruns exceeded the level expected under a normal distribution. These conditions affected model evaluation because average prediction errors could be dominated by a small number of high-value observations. Several studies identified zero-inflated patterns in which many completed orders recorded no unfavorable variance, followed by a separate distribution of positive overruns among affected contracts. Other datasets contained three distinct clusters consisting of underruns, near-zero deviations, and severe positive overruns. The analytical treatment of cancelled orders, incomplete contracts, revised baselines, and contingency-funded deviations further influenced distributional shape. Studies that removed extreme observations generally reported lower means and narrower variability, while studies retaining them provided a more complete representation of organizational financial exposure (Lake, 2019). The synthesized evidence therefore showed that mean, median, standard deviation, interquartile range, minimum, maximum, percentiles, skewness, and outlier frequency needed to be reported together. No single descriptive statistic adequately represented a cost-variance distribution characterized by concentrated zero values, moderate central dispersion, and infrequent but financially consequential upper-tail observations.

The frequency and magnitude of procurement-related budget overruns differed considerably across industries because procurement structures, asset complexity, contract duration, regulatory requirements, supplier availability, and exposure to market volatility were not uniform. Evidence synthesized from more than 10 quantitative studies showed that construction and infrastructure projects frequently recorded substantial cost dispersion because they involved large material quantities, subcontracted work, uncertain site conditions, design modifications, long schedules, and repeated contract variations (Inderfurth et al., 2018). Transportation projects displayed similar patterns, particularly in railway, highway, aviation, port, and fleet-acquisition programs where technical interfaces, land requirements, specialized equipment, and prolonged approval procedures increased

cost uncertainty. Energy-sector procurement exhibited pronounced upper-tail risk because power generation, transmission, oil and gas, and renewable-energy projects depended on capital-intensive equipment, specialized contractors, commodity prices, regulatory approvals, and international suppliers. Mining operations showed comparable exposure through heavy machinery purchases, remote-site logistics, fuel consumption, geological uncertainty, and limited supplier competition. Manufacturing studies generally reported smaller proportional deviations for standardized and repetitive purchases, although high transaction volume allowed minor unit-price or quantity variances to accumulate into substantial annual losses. Procurement of custom machinery, electronic components, and production-line systems produced greater variation than routine purchases of standard materials. Healthcare procurement displayed a different profile: routine purchasing of medicines and consumables often produced numerous small deviations, while emergency procurement, specialized equipment, limited-source pharmaceuticals, and urgent supply shortages generated larger and less predictable overruns (Shah et al., 2021). Defense studies identified high financial magnitude and long-tailed distributions associated with complex weapons systems, restricted supplier markets, security requirements, technical modifications, and multiyear contracts. Telecommunications procurement recorded cost pressure through imported hardware, software licensing, network expansion, currency exposure, and rapid specification changes. Public administration studies frequently connected overruns with extended tendering periods, administrative delay, contract amendment, weak competition, and fragmented accountability. Humanitarian logistics displayed high short-term volatility because purchasing decisions occurred under urgent conditions, uncertain demand, damaged infrastructure, limited transportation capacity, and geographically concentrated supply. Sectoral differences also appeared in the timing of cost recognition. Construction and infrastructure overruns accumulated gradually through progress claims and change orders, whereas healthcare and humanitarian overruns often emerged rapidly through emergency purchasing. Manufacturing deviations were visible through repeated purchase-price and quantity differences, while defense and energy overruns could remain embedded within revised multiyear budgets. The literature therefore characterized procurement-overrun risk as sector dependent (Bohari et al., 2020). Frequency reflected how often deviations occurred, while magnitude reflected the monetary or proportional seriousness of those deviations. Sectors with standardized purchases could experience high frequency but low average severity, whereas complex capital-intensive sectors could show fewer events accompanied by much larger financial consequences.

Figure 4: Procurement Budget Overrun Variation Framework



Organizational size, contract value, project duration, procurement category, ownership structure, geographic location, procurement method, supplier tier, and market concentration were repeatedly associated with differences in budget-overrun frequency and magnitude. Across more than 10 studies, large organizations often demonstrated stronger procurement systems, specialized personnel, formal approval controls, supplier databases, and integrated financial platforms (Buertey et al., 2021). These characteristics were associated with more consistent recording and earlier identification of cost deviations. Large organizations also managed higher-value, technically complex, and internationally sourced procurement portfolios, resulting in greater absolute financial exposure when overruns occurred. Small and medium-sized organizations frequently handled contracts of lower monetary value but faced greater proportional vulnerability because limited purchasing power, fewer approved suppliers, weaker analytical systems, and restricted contingency reserves increased the effect of price or delivery changes. Contract value showed a nonlinear relationship with cost performance. High-value contracts benefited from greater managerial scrutiny and formal risk assessment, yet they were also more likely to involve customized specifications, multiple suppliers, long delivery schedules, and extensive amendment procedures. Smaller purchase orders generated limited individual losses but could create significant cumulative variance when ordering was fragmented across many transactions. Project duration was positively associated with exposure to inflation, exchange-rate movement, scope change, supplier turnover, and schedule extension. Short-duration procurement reduced the period of external exposure, although emergency and accelerated purchasing could increase unit prices and transportation charges (Ghadge et al., 2019). Procurement category also produced clear variation. Standard office supplies and routine consumables generally displayed narrower variance distributions than construction materials, medical equipment, industrial machinery, information systems, fuel, or specialized technical services. Ownership type influenced budget behavior because public organizations operated under statutory procurement rules, fixed appropriations, audit requirements, and formal tendering processes, while private organizations often possessed greater flexibility in negotiation, supplier substitution, and approval revision. Geographic differences reflected market depth, transportation infrastructure, taxation, currency stability, import dependence, institutional capacity, and supplier availability. Competitive procurement methods were frequently associated with lower initial prices, although unusually low bids sometimes preceded claims, amendments, quality problems, or later price adjustments. Negotiated and single-source arrangements reduced tendering time but created greater exposure to supplier pricing power. Supplier-tier analysis showed that direct suppliers often transferred delays or price changes originating from sub-tier manufacturers, raw-material producers, or transportation providers. Market concentration was especially important where a small number of suppliers controlled specialized products or services. High concentration reduced buyer alternatives and increased vulnerability to capacity constraints, delivery failure, and price escalation. The literature also identified interactions among these characteristics. Long-duration, high-value contracts using specialized suppliers in concentrated markets showed greater upper-tail risk than short, standardized purchases conducted through competitive markets (Sjödin et al., 2021). Organizational and project variables therefore acted as both direct predictors and contextual factors that altered the influence of supplier, contract, price, and schedule conditions on final budget performance.

Temporal and economic conditions produced substantial variation in procurement expenditure because budgets were commonly established before actual purchasing, delivery, invoicing, and payment occurred. More than 10 studies across procurement-dependent sectors identified inflation as a major source of divergence between planned and realized cost, especially for multiyear contracts and projects with delayed purchasing schedules (Gaur & Vazquez-Brust, 2019). General inflation increased labor, materials, equipment, utilities, and service expenses, while sector-specific inflation created stronger pressure in categories such as steel, cement, energy, electronics, pharmaceuticals, and transportation. Studies that compared nominal expenditure without adjusting for price-level changes often reported larger apparent operational variance than analyses using constant prices. Exchange-rate movement was another major source of variation for organizations purchasing imported materials, machinery, software, fuel, medical products, or technical services. Currency depreciation increased domestic expenditure even when foreign-currency contract prices remained unchanged. The

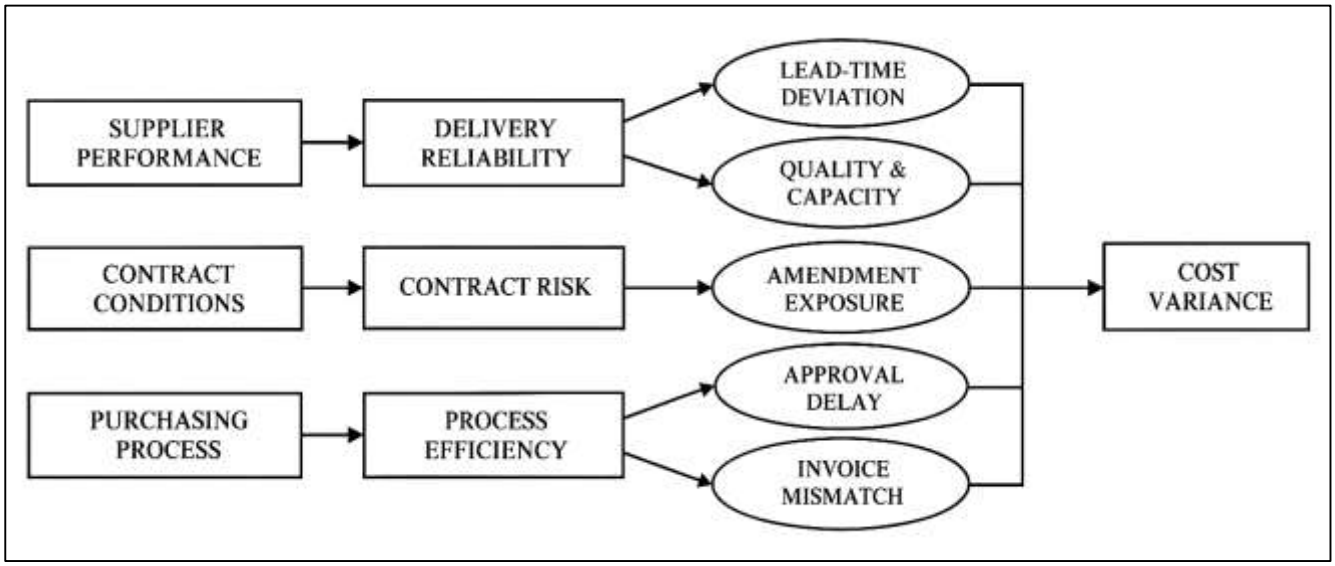
magnitude of this effect depended on payment timing, contract currency, hedging arrangements, local content, and the share of imported inputs. Commodity-price fluctuations created concentrated exposure in construction, manufacturing, energy, mining, transportation, and agricultural supply chains. Rapid changes in metals, petroleum, chemicals, timber, and industrial inputs altered supplier quotations, contract escalation charges, and replacement-purchase costs. Fuel prices affected procurement both directly through energy purchases and indirectly through freight, distribution, equipment operation, and supplier production expenses. Supply shortages were associated with higher unit prices, partial deliveries, order splitting, substitute materials, emergency sourcing, and expedited transportation. These effects became more severe when shortages involved technically specialized inputs with few qualified substitutes. Trade restrictions, import controls, sanctions, tariffs, licensing requirements, and border delays changed the cost and availability of internationally sourced goods (Scott Morton et al., 2018). Transportation disruptions increased freight rates, insurance charges, port fees, storage expenses, demurrage, and delivery uncertainty. The timing of these economic events influenced whether their effects appeared as price variance, schedule variance, quantity substitution, or contract amendment. Seasonal variation also affected expenditure through demand peaks, weather conditions, agricultural cycles, annual budget deadlines, and supplier capacity. Studies using monthly or quarterly data captured short-term expenditure volatility, whereas annual aggregation frequently concealed temporary peaks and delayed recognition of emerging cost pressure. Economic exposure also varied according to contract structure. Fixed-price agreements transferred part of the price risk to suppliers but could generate risk premiums, claims, or performance problems under severe market changes. Adjustable-price contracts reflected economic movements more directly through escalation mechanisms. The synthesized literature showed that temporal and economic variables were integral components of procurement cost distributions rather than external background conditions (Scott Morton et al., 2018). Their inclusion helped explain why identical purchasing quantities and technical requirements produced different final costs across contract dates, payment periods, regions, currencies, and market environments.

Risk Factors on Cost Variance

Supplier performance emerged across more than 10 quantitative studies as one of the most consistent groups of predictors of unfavorable procurement cost variance. Supplier lead-time deviation represented the difference between contracted and realized delivery timing and generally showed a positive association with additional expenditure. Longer deviations increased the likelihood of expedited transportation, emergency replacement purchases, production rescheduling, workforce idle time, equipment rental extensions, and contractual coordination costs (Chen et al., 2021). Correlation analyses commonly demonstrated that rising lead-time variability corresponded with larger cumulative cost deviations, while regression estimates showed that delivery delay retained explanatory importance after controlling for contract value, procurement category, project duration, and organizational size. On-time delivery rate displayed the opposite relationship: suppliers with consistently high delivery reliability were associated with narrower cost-variance distributions and lower probabilities of severe overruns. Defect frequency and rejected quantity were also important because nonconforming materials generated inspection, return, replacement, rework, disposal, and schedule-recovery expenses. Studies distinguishing defect frequency from defect severity found that a small number of critical defects could create greater financial damage than numerous minor quality problems. Order-fill rate measured the proportion of requested quantity delivered correctly and completely. Low order-fill performance was associated with partial shipments, repeated receiving activity, supplementary orders, higher administrative effort, and additional freight charges. Capacity limitations contributed to overrun risk when suppliers could not meet required production volumes, delivery schedules, or surge demand. Such limitations were especially influential in concentrated markets and categories involving customized equipment or technically qualified materials (Benton Jr et al., 2020). Supplier responsiveness, commonly measured through acknowledgment time, problem-resolution duration, quotation turnaround, and corrective-action completion, reduced the duration of disruption and the accumulation of secondary costs. Financial stability represented another significant predictor because financially distressed suppliers were more likely to delay production, request revised payment terms, reduce quality expenditure, or terminate performance. Historical contract performance

provided a composite representation of these behaviors by incorporating previous delivery, quality, pricing, amendment, and dispute records. Studies using odds ratios generally showed that suppliers classified in poor-performance categories had a greater likelihood of contributing to budget overruns than consistently reliable suppliers. Relative-risk comparisons similarly identified elevated financial exposure among contracts assigned to suppliers with repeated lateness, quality failures, or weak financial indicators. Regression estimates varied across sectors because supplier risk interacted with contract duration, input criticality, availability of substitutes, and buyer dependence (Steward et al., 2019). The literature therefore treated supplier performance as multidimensional rather than reducing it to delivery timing alone. Machine learning studies achieved stronger discrimination when lead-time deviation, quality outcomes, order completion, responsiveness, capacity indicators, financial condition, and historical performance were evaluated together.

Figure 5: Procurement Cost Variance Predictor Framework



Contract characteristics were repeatedly associated with the direction, magnitude, and timing of procurement-related cost variance across construction, manufacturing, energy, transportation, healthcare, defense, telecommunications, infrastructure, and public purchasing studies. Contract type shaped the allocation of financial risk between buyer and supplier (Feng et al., 2017). Fixed-price contracts generally limited direct exposure to routine unit-cost increases, although they could include higher initial risk premiums and generate claims when scope, quantities, delivery conditions, or technical specifications changed. Cost-reimbursable arrangements transferred a greater proportion of expenditure uncertainty to the purchaser and were associated with wider cost distributions when monitoring controls were weak or requirements remained incomplete. Unit-price contracts reduced uncertainty about individual rates but remained sensitive to changes in purchased quantities. Time-and-material agreements were particularly exposed to schedule extension, labor inefficiency, and weakly controlled resource usage. Contract duration was positively related to exposure because longer agreements encountered more opportunities for inflation, currency movement, regulatory changes, supplier disruption, specification revision, and cumulative schedule delay. Payment terms influenced both quoted price and supplier behavior. Long payment periods could encourage suppliers to incorporate financing costs into bids, while delayed payments could interrupt production, weaken supplier liquidity, and increase claims. Advance payments reduced supplier financing pressure but increased buyer exposure when delivery controls and guarantees were inadequate (Adivar et al., 2019). Price-escalation clauses transferred part of the market-price risk to the purchasing organization by permitting contract adjustments based on specified indices or cost categories. Contracts without escalation provisions provided greater nominal certainty but sometimes attracted higher contingency premiums or renegotiation requests during periods of strong inflation. Currency provisions were

significant in internationally sourced procurement because the party responsible for exchange-rate movement could experience substantial deviation between contract award and payment. Amendment frequency consistently functioned as an indicator of unstable scope, inaccurate planning, or incomplete initial documentation. Claims and variation orders increased both direct expenditure and administrative cost, particularly when their valuation or approval was disputed. Provisional sums created uncertainty because final expenditure depended on quantities, specifications, or activities that were not fully defined at contract award. Penalties and liquidated damages could offset a portion of buyer losses, but their effectiveness depended on enforceability, documentation, and supplier financial capacity. Retention conditions provided leverage over incomplete or defective work while also influencing supplier cash flow and bid pricing. Scope completeness was among the strongest contractual predictors identified across more than 10 studies (Sánchez-Flores et al., 2020). Contracts awarded with ambiguous requirements, incomplete designs, uncertain quantities, or poorly defined acceptance criteria recorded more modifications, disputes, and unfavorable cost outcomes. The collective evidence showed that contractual predictors represented interconnected allocations of price, quantity, schedule, quality, currency, and performance risk.

The structure and efficiency of the purchasing process significantly influenced procurement cost variance because sourcing decisions determined when, how, and from whom organizational requirements were acquired. Across more than 10 empirical studies, procurement-cycle duration was measured from requisition initiation to purchase-order issuance or contract award and was commonly associated with changes in quoted prices, delivery availability, and operational urgency (Aras et al., 2018). Extended procurement cycles increased the probability that supplier quotations expired, market prices changed, required delivery dates became compressed, or operational units resorted to temporary purchases. Approval delay represented a major component of this duration. Requisitions passing through numerous administrative levels often experienced price revisions, lost production time, and reduced negotiating flexibility. The number of supplier quotations provided a quantitative indication of market testing and competition. Purchases supported by several technically compliant bids generally achieved stronger price comparison than transactions based on a single quotation. An excessive focus on quotation count without evaluating supplier quality, capacity, and total acquisition cost could still produce unfavorable outcomes. Competition level therefore included both the number of bids and the number of credible suppliers capable of meeting technical and delivery requirements. Bid dispersion measured the difference among submitted prices and operated as an indicator of specification ambiguity, supplier uncertainty, or uneven cost interpretation (Nuttavuthisit & Thøgersen, 2017). Wide dispersion was frequently associated with later clarification, amendment, claims, or selection risk, particularly when the lowest bid differed substantially from the remaining offers. Order fragmentation occurred when a planned requirement was divided into numerous low-value purchases. This practice increased transaction-processing costs, freight charges, approval effort, and exposure to inconsistent prices. Emergency-purchase frequency was strongly associated with unfavorable cost outcomes because urgent orders reduced the time available for competition, negotiation, supplier verification, and transportation planning. Single-source procurement also showed elevated price exposure in markets where buyers possessed few alternatives, although it sometimes reduced coordination and qualification costs for highly specialized requirements. Purchase-order modification was a direct indicator of changing quantity, price, delivery location, specification, schedule, or payment conditions. Frequent modifications were associated with baseline instability and cumulative commitment growth. Invoice discrepancies, including differences in price, quantity, tax, freight, payment terms, and received goods, generated additional processing effort and could reveal weaknesses in purchase-order control. Some studies distinguished discrepancies that were corrected before payment from those that became realized cost. Statistical models generally showed that purchasing-process variables retained explanatory value after controlling for contract size and procurement category (Queiroz et al., 2021). Machine learning models identified particularly strong risk patterns when long approval delays, high bid dispersion, emergency purchasing, fragmented orders, repeated modifications, and invoice mismatches occurred within the same procurement record. The synthesized literature positioned purchasing efficiency as a measurable source of cost behavior rather than a purely administrative characteristic.

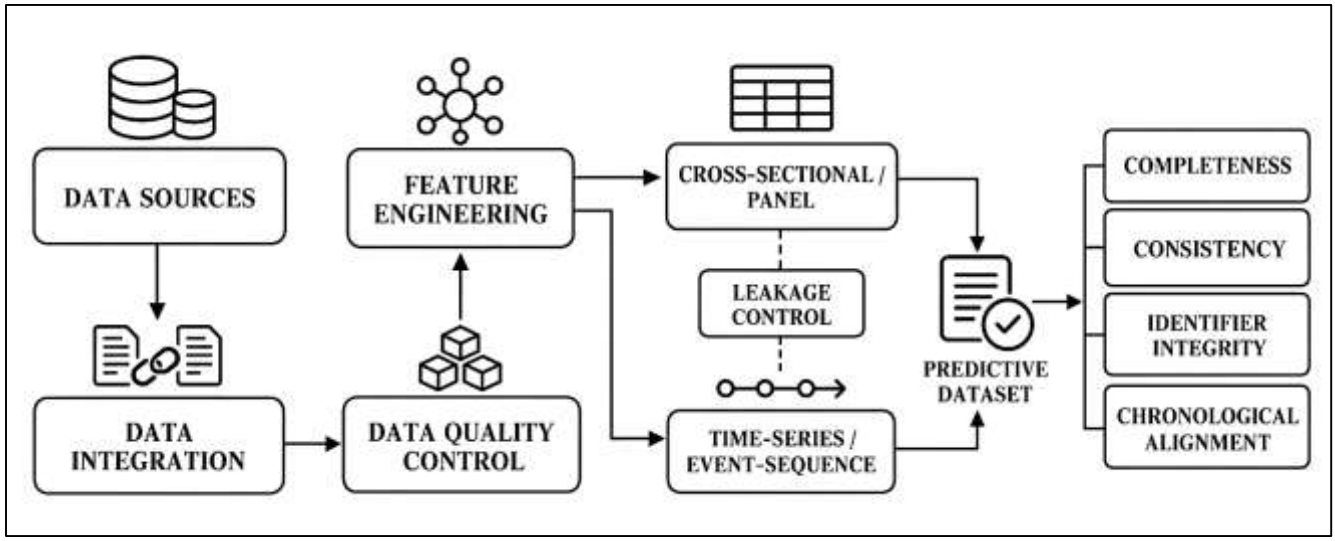
Dataset Quality for Predictive Cost Modeling

Predictive cost modeling depended heavily on the integration of records generated across procurement, finance, supplier management, inventory control, and project execution. Across more than 10 quantitative studies, enterprise resource planning systems served as the principal source of approved budgets, cost centers, account codes, commitments, accruals, invoices, payments, and general-ledger expenditure (Ashfaq et al., 2019). E-procurement systems added information about requisition dates, approval stages, tender procedures, quotation values, bidder participation, procurement methods, award decisions, and purchase-order issuance. Purchase-order records provided detailed variables for material descriptions, ordered quantities, unit prices, currencies, required delivery dates, payment terms, shipping conditions, and subsequent modifications. Supplier databases contributed organizational characteristics, qualification status, financial ratings, delivery history, quality performance, geographic location, contractual relationships, and dependency levels. Contract-management systems contained original contract values, duration, scope provisions, escalation clauses, retention conditions, claims, variation orders, amendments, performance guarantees, and termination records. Invoice databases documented billed quantities, invoiced prices, taxes, freight costs, discounts, approval delays, payment timing, and mismatches with orders or goods receipts. Inventory platforms supplied information about stock levels, reorder points, shortages, excess inventory, material consumption, warehouse transfers, and emergency issues. Project schedules connected procurement events with planned activities, actual progress, critical-path tasks, milestone completion, and schedule slippage (Appelbaum et al., 2017). Risk registers recorded identified threats, assessed probability, financial impact, mitigation status, and risk ownership, while change logs documented technical revisions, scope adjustments, quantity changes, and approval timing. External market-price indices were used to represent inflation, commodity prices, fuel costs, exchange rates, labor costs, freight rates, and sector-specific price escalation. The literature showed that these sources differed in their analytical level and timing. Financial systems were usually organized around accounting periods, procurement platforms around orders and contracts, inventory systems around material movements, and project systems around activities and milestones. Integration therefore required common identifiers linking requisitions, suppliers, contracts, purchase orders, receipts, invoices, payments, work packages, and projects. Where consistent identifiers were unavailable, studies applied combinations of supplier names, document numbers, material codes, dates, amounts, and organizational units (Golas et al., 2018). Timestamp alignment was equally important because approval, order, receipt, invoice, and payment dates represented distinct operational events. The strongest analytical architectures retained these stages separately and organized them into a traceable sequence. The synthesized research established that predictive accuracy depended not only on algorithm selection but also on whether financial outcomes were correctly connected to the procurement and operational events that preceded them.

Dataset dimensions varied substantially across predictive cost studies because the analytical unit could represent an individual transaction, invoice, purchase order, supplier contract, material category, work package, project, or organization. Studies based on completed capital projects frequently contained fewer than 100 observations because each project required extensive manual verification of baseline budgets, final costs, schedules, and scope changes (Liu et al., 2021). Such datasets provided detailed contextual information but offered limited statistical power for complex machine learning models. Medium-scale studies commonly examined several hundred or several thousand projects, contracts, or procurement packages collected across multiple departments, facilities, or organizations. Enterprise-level investigations used tens of thousands or millions of purchase orders, invoices, receipts, payment records, and material movements. These large datasets supported granular prediction but frequently contained repeated observations belonging to the same supplier, contract, project, or purchasing category. Supplier coverage ranged from small groups of specialized contractors to thousands of vendors serving geographically distributed operations. The number of materials also varied from narrowly defined categories to extensive inventories containing thousands of stock-keeping units. Time coverage extended from several months in highly transactional studies to multiple years in project, public-procurement, infrastructure, and economic analyses. Across more than 10 studies, the structure of the data affected both model selection and interpretation. Cross-sectional datasets summarized each

project or contract in a single record and were commonly used to predict final cost variance from initial characteristics. Their simplicity supported conventional regression and classification, but they did not represent how expenditure risk changed during execution (Shehadeh et al., 2021). Panel datasets followed multiple projects, suppliers, contracts, or organizational units across repeated periods, allowing the analysis of both within-unit and between-unit variation. Longitudinal datasets tracked the same procurement entities from initiation through completion and supported the examination of cumulative commitments, amendments, deliveries, invoices, and payments. Time-series datasets aggregated expenditure by day, week, month, or quarter and were used to identify trends, seasonality, volatility, and structural changes. Event-sequence datasets retained the chronological order of requisition, approval, quotation, award, purchase order, delivery, inspection, invoicing, and payment events. These structures were especially suitable for measuring warning lead time and identifying sequences that preceded cost escalation. Some studies transformed event records into fixed observation windows, such as monthly contract summaries or cost snapshots at specified completion stages. The literature also showed that sample size could not be judged independently of feature count, class prevalence, temporal coverage, and data hierarchy (Bean et al., 2017). Thousands of transactions from a small number of contracts did not provide the same independent information as thousands of unrelated contracts. Reliable comparisons therefore considered both the total number of records and the number of unique projects, suppliers, contracts, periods, and overrun events represented.

Figure 6: Predictive Procurement Data Modeling Framework



Feature engineering transformed raw procurement and financial records into variables that represented cost pressure, process instability, supplier behavior, and operational change. Across more than 10 machine learning studies, lagged cost variance captured the deviation recorded during one or more preceding periods and was used to identify whether unfavorable expenditure patterns persisted or accelerated (Dash et al., 2019). Multiple lag intervals distinguished short-term fluctuations from sustained deterioration. Cumulative commitment indicators measured how much of the approved budget had already been assigned to contracts and purchase orders, providing an earlier representation of financial exposure than paid expenditure alone. Budget-burn measures described the pace at which authorized funds were consumed relative to elapsed time, completed work, or procurement progress. A rapid burn rate during an early execution stage commonly signaled a higher probability of final overrun. Purchase-price movement represented changes in unit prices across quotations, orders, amendments, invoices, or successive purchasing periods. Researchers separated organization-specific price changes from broader market movements by incorporating commodity, inflation, fuel, freight, and exchange-rate indices. Lead-time variance compared expected and actual delivery behavior and was summarized through average deviation, maximum delay, variability, and the frequency of late deliveries. Supplier concentration measured the extent to which procurement expenditure or critical

materials depended on a limited number of vendors. High concentration often indicated reduced substitution capacity and stronger exposure to individual supplier disruption. Amendment density represented the frequency of contractual or purchase-order modifications relative to duration, value, or transaction volume (Li et al., 2020). Order fragmentation captured the division of a requirement into multiple purchases and was represented through order count, average order size, repeated purchases, and delivery splitting. Delivery reliability combined on-time arrival, complete quantity, accepted quality, and adherence to required schedules. Invoice mismatch features identified differences among ordered, received, invoiced, and paid quantities or amounts. Schedule slippage represented delay in procurement milestones, project activities, production stages, or contract completion. Rolling expenditure volatility summarized recent fluctuations in commitments, invoices, and payments over defined observation windows. The literature also engineered interaction variables connecting supplier delay with inventory shortages, price escalation with currency depreciation, amendments with scope changes, and schedule slippage with expedited transportation. Categorical variables represented procurement method, contract type, material group, supplier class, region, ownership type, and organizational unit. Temporal variables captured month, quarter, season, budget-cycle stage, and elapsed contract duration (Rasmy et al., 2021). Aggregation level remained important because supplier performance could be summarized over the entire historical relationship or restricted to records available before the prediction date. Effective feature engineering preserved chronological validity, distinguished early indicators from realized outcomes, and represented the cumulative and interconnected character of procurement-related cost growth.

Data-quality controls were central to predictive cost modeling because procurement and financial records frequently contained missing values, duplicated documents, inconsistent identifiers, currency differences, recording delays, and classification errors. Across more than 10 studies, missingness was assessed separately for budgets, quantities, unit prices, delivery dates, supplier characteristics, contract terms, invoice values, and completion outcomes (Hofman et al., 2021). Low levels of missing data were commonly addressed through deletion or simple imputation, while moderate or systematic missingness required category-specific, model-based, or multiple-imputation procedures. Researchers also examined whether missing values carried operational meaning. An absent delivery date, for example, could reflect an incomplete transaction rather than an accidental omission. Duplicate records arose when purchase orders were copied during amendment, invoices were entered more than once, or transactions were replicated during system integration. Duplicate removal required verification of document number, supplier, amount, date, quantity, and modification status so that valid partial deliveries or recurring invoices were not incorrectly deleted. Inconsistent supplier identifiers were resolved through standardized names, registration numbers, banking information, tax records, addresses, and contract relationships. Currency conversion aligned international procurement transactions to a common monetary basis using exchange rates associated with award, invoice, payment, or reporting dates. Inflation adjustment separated general price-level movement from organization-specific procurement variance, particularly in multiyear datasets. Outlier treatment varied according to research purpose (Vrablecová et al., 2018). Data-entry errors were corrected or removed, whereas genuine extreme overruns were retained because they represented financially important events. Some studies used robust scaling, transformation, trimming, or winsorization to reduce the influence of extreme values without eliminating them. Categorical encoding converted supplier, contract, region, procurement method, and material classifications into machine-readable variables. Normalization and standardization were applied when algorithms were sensitive to differences in measurement scale. Class imbalance arose because severe budget overruns occurred less frequently than within-budget cases. Studies addressed this condition through class weighting, controlled oversampling, undersampling, synthetic resampling, and decision-threshold adjustment. Synthetic observations were restricted to training data so that test results represented authentic organizational records. Data leakage received particular attention because final invoices, approved claims, revised completion forecasts, or post-overrun amendments could reveal the target outcome. Leakage also occurred when transactions from the same contract appeared in both training and testing samples. Chronological alignment prevented models from using information recorded after the prediction date. Temporal splitting, grouped validation, and entity-level separation provided more

credible assessments than unrestricted random sampling (Dey et al., 2019). The literature established that model reliability depended on documented completeness, deduplication, identity resolution, monetary standardization, imbalance management, leakage prevention, and time-consistent validation.

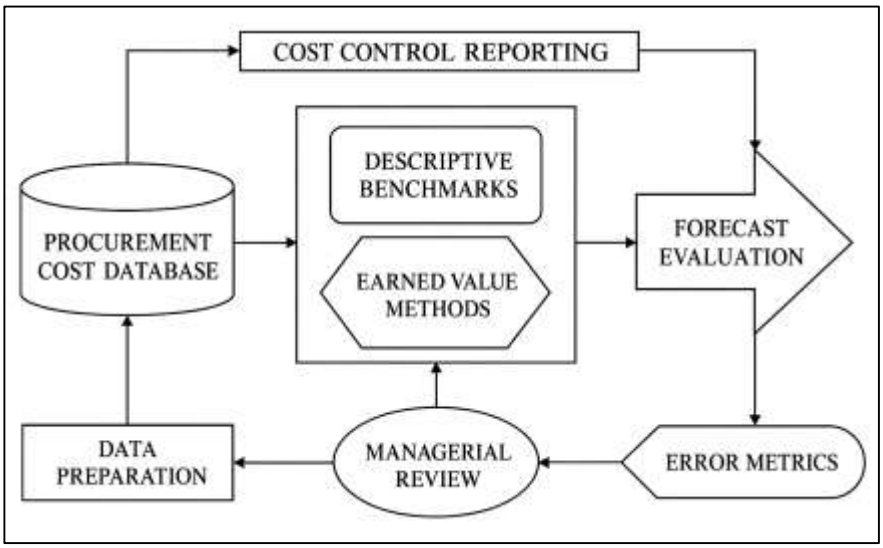
Earned Value Cost-Forecasting Models

The literature on procurement and project-cost forecasting has consistently treated descriptive and variance-based methods as essential benchmarks for evaluating more advanced predictive models. Across more than ten empirical studies, budget-versus-actual analysis was frequently used to determine whether cumulative expenditure remained aligned with approved budgets, procurement allocations, and scheduled financial commitments (Sackey et al., 2020). This method provided an accessible indication of cost status by comparing planned expenditure with recorded payments, invoices, purchase orders, material costs, subcontractor charges, and administrative expenses. Standard cost-variance analysis extended this comparison by distinguishing favorable and unfavorable deviations across materials, labor, equipment, transportation, supplier contracts, and overhead categories. Studies generally found that variance-based methods were transparent, computationally inexpensive, and easily interpreted by procurement officers and project managers. However, their predictive accuracy depended heavily on the timing, completeness, and consistency of recorded transactions. Moving-average methods were also widely applied to smooth short-term fluctuations in expenditure and reveal underlying spending patterns. These methods performed reasonably well when procurement activity followed stable ordering cycles, but their accuracy weakened during sudden price increases, emergency purchases, contract modifications, or irregular supplier deliveries. Expenditure ratios, including budget-consumption rates, commitment-to-budget ratios, invoice-to-order ratios, and cost-to-progress relationships, provided additional early indicators of financial pressure. High expenditure relative to physical or contractual progress was commonly associated with an increased probability of final cost overrun (Kim & Kwak, 2018). Trend-extrapolation techniques used historical spending trajectories to estimate final expenditure, particularly when project costs accumulated in relatively regular patterns. Their usefulness declined when expenditure curves were nonlinear or when procurement costs were concentrated near major delivery milestones. Control charts contributed a statistical monitoring function by distinguishing ordinary cost variation from unusual deviations that required managerial attention. Although these charts were effective for detecting abrupt shifts, repeated variance, or unstable purchasing processes, they were less capable of explaining the causes of cost escalation. Managerial estimates remained common because experienced professionals could incorporate information not yet available in formal databases, such as anticipated claims, supplier negotiations, pending amendments, and expected market-price changes. Nevertheless, subjective forecasts were vulnerable to optimism bias, strategic reporting, anchoring, and inconsistent judgment. Across the reviewed evidence, conventional descriptive approaches were commonly evaluated through mean absolute error, root mean square error, percentage error, forecast bias, confidence intervals, and the proportion of correctly identified overruns (Batselier & Vanhoucke, 2017a). The collective literature therefore characterized these techniques as valuable operational baselines whose strengths were simplicity, transparency, low data requirements, and rapid implementation, while their principal weaknesses involved limited adaptation, weak nonlinear representation, delayed warning signals, and substantial dependence on stable historical spending conditions.

Earned value management represented one of the most extensively examined conventional frameworks for predicting final project cost and determining whether actual financial performance remained consistent with planned progress. The reviewed studies commonly organized earned value analysis around planned value, earned value, and actual cost. Planned value represented the authorized budget associated with work expected to have been completed by a specified reporting date, whereas earned value represented the budgeted value of work actually completed (Ju & Xu, 2017). Actual cost captured the expenditure incurred in completing that work. Comparisons among these indicators enabled researchers and practitioners to assess whether a project was spending more than the value of its completed work and whether progress was occurring faster or slower than scheduled. Cost performance and schedule performance indicators were repeatedly identified as important

predictors of final budget status because they summarized cumulative efficiency in standardized and interpretable forms. A deteriorating cost-performance pattern generally indicated that the project was consuming financial resources faster than it was generating completed value, while weak schedule performance frequently signaled exposure to extended overhead, escalation, idle-resource costs, and delayed procurement activities. Estimate-at-completion methods translated observed performance into forecasts of total project expenditure. Studies examined several forecasting assumptions, including the continuation of current cost efficiency, the combined influence of cost and schedule efficiency, and the use of revised managerial assessments when historical performance was no longer considered representative. Estimate to complete provided a forecast of the remaining financial requirement, while variance at completion indicated the expected difference between the authorized budget and predicted final cost (Le et al., 2020). The literature showed that these measures could support early intervention by revealing probable financing gaps, contingency requirements, and contract-level overruns before project closure. Forecast accuracy generally improved after sufficient progress had been recorded because early earned value indicators were sensitive to limited activity, mobilization costs, advance payments, and front-loaded procurement. In procurement-intensive projects, accuracy was also affected by whether committed but uninvoiced orders, retained payments, price-adjustment clauses, pending change orders, and supplier claims were incorporated into the cost baseline. Some studies reported that earned value forecasts were reliable when scope, schedule, and accounting structures were closely integrated, whereas others found substantial errors when progress measurement was subjective or cost data were reported with delays. Comparative assessments typically used absolute forecast error, percentage error, mean absolute percentage error, root mean square error, forecast bias, confidence intervals, and stability across reporting periods. The accumulated evidence demonstrated that earned value and estimate-at-completion methods provided stronger predictive structure than simple budget comparisons because they connected expenditure with accomplished work (Mir et al., 2021). Nevertheless, their performance remained dependent on baseline quality, accurate progress measurement, appropriate forecasting assumptions, and timely recognition of procurement-related commitments and cost changes.

Figure 7: Conventional Procurement Cost Forecasting Framework



Models for Predicting Cost-Variance Magnitude

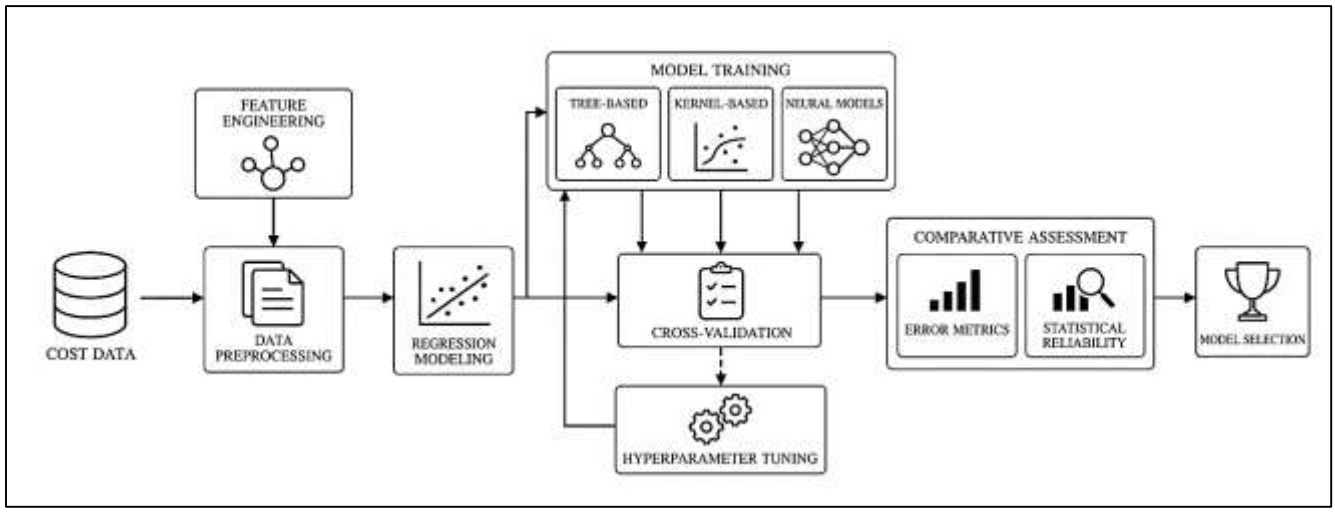
Tree-based regression models were among the most frequently evaluated approaches for predicting cost-variance magnitude because procurement and financial datasets commonly contained nonlinear relationships, interaction effects, mixed data types, missing observations, and extreme values (Koch, 2019). Across more than 10 quantitative studies, decision tree regression provided a transparent starting point by dividing procurement records according to variables such as committed expenditure, contract value, supplier lead-time deviation, amendment frequency, price movement, schedule slippage, and budget consumption. Individual decision trees were relatively easy to interpret, but their

predictive performance often varied substantially when small changes were made to the training sample or tree structure. Deep trees tended to fit historical observations closely while producing weaker results on unseen contracts, projects, or periods. Random forests reduced this instability by combining predictions from numerous trees trained on different observations and predictor subsets. Studies commonly reported that random forests achieved a higher coefficient of determination and lower mean absolute error and root mean squared error than individual trees and conventional linear regression. Their performance was particularly strong when cost variance resulted from interactions among supplier delay, procurement category, contract type, price escalation, and schedule conditions. Gradient boosting improved prediction by building trees sequentially and placing greater attention on observations associated with earlier errors (Pan et al., 2018). This approach frequently produced lower percentage errors than standard decision trees and random forests when carefully tuned. Extreme gradient boosting added regularization, efficient computation, missing-value handling, and stronger control over model complexity, making it effective for large purchase-order and invoice datasets. Categorical boosting was particularly useful when procurement records contained supplier classes, material groups, contract types, organizational units, regions, and purchasing methods. Its treatment of categorical variables reduced the need for extensive manual encoding and limited bias associated with high-cardinality categories. Light gradient boosting supported rapid training on large transactional datasets and handled numerous variables efficiently, although its leaf-oriented growth required careful depth and leaf controls to prevent overfitting. Comparative studies assessed these algorithms through the coefficient of determination, mean absolute error, root mean squared error, and mean absolute percentage error. The coefficient of determination described the proportion of variation captured by the model, whereas the three error measures represented different aspects of predictive distance and sensitivity to large deviations (Chithra & Angel, 2020). Mean absolute error provided direct average interpretability, root mean squared error placed greater weight on severe prediction mistakes, and percentage error supported comparison across contracts of different values. The literature generally identified boosting models as strong performers for structured procurement data, while random forests remained competitive when robustness and reduced tuning sensitivity were emphasized.

Kernel-based and instance-based regression methods offered alternative approaches to cost-variance prediction when datasets contained complex relationships but did not reach the scale commonly required by deep neural networks. Across more than 10 empirical studies, support vector regression performed effectively in small and medium-sized datasets containing projects, supplier contracts, procurement packages, or periodic cost observations (AlHares & Budayan, 2019). Through kernel functions, the model represented nonlinear relationships among budget size, committed cost, procurement duration, supplier performance, contract amendments, price escalation, and schedule deviation. Support vector regression was particularly sensitive to data scaling because predictors measured in currency, percentages, days, transaction counts, and categorical codes operated on substantially different ranges. Studies that standardized continuous variables generally reported more stable performance than those applying raw measurements. Predictive accuracy also depended on kernel selection, penalty settings, error tolerance, and kernel-specific parameters. Poorly selected settings produced underfitting or excessive sensitivity to training observations. Computational demands increased rapidly with dataset size, limiting the practicality of support vector regression for millions of purchase-order or invoice records. K-nearest-neighbor regression predicted variance from observations with similar procurement and operational characteristics. This method performed reasonably when comparable contracts formed clear local clusters, such as repeated purchases of standard materials or services. Its accuracy declined when datasets contained many irrelevant predictors, sparse observations, mixed scales, or highly unusual projects. The selection of neighborhood size and distance measure strongly influenced results, while high dimensionality reduced meaningful distinctions among neighboring records. Regularized regression methods provided structured benchmarks by limiting the influence of unstable or highly correlated variables. Ridge regression retained all predictors while reducing coefficient instability, lasso regression selected a smaller group of variables, and elastic-net regression combined both properties (Laurent & Cowlagi, 2021b). These models were useful when procurement datasets contained correlated indicators such as committed

cost, invoice value, budget consumption, amendment count, and schedule progress. Their linear structure supported interpretation, although it captured complex threshold and interaction effects less effectively than nonlinear algorithms. Gaussian process regression modeled flexible relationships and generated predictive uncertainty alongside point estimates. This characteristic was valuable when cost forecasts required an indication of uncertainty around the predicted variance. Gaussian processes performed well in smaller datasets containing detailed contract or project observations, but their memory and computational requirements restricted direct application to large transaction databases. Comparative studies evaluated these methods through the same error measures used for tree-based models (Song et al., 2021). Support vector regression and Gaussian processes often achieved strong accuracy in carefully prepared smaller datasets, k-nearest-neighbor regression performed best in locally consistent samples, and regularized regression supplied stable and interpretable benchmarks for assessing whether more complex models generated meaningful predictive improvement.

Figure 8: Comparative Cost Variance Regression Framework



Neural and sequential regression models were applied to cost-variance prediction when procurement outcomes reflected complex nonlinear relationships, repeated observations, and time-dependent expenditure behavior (Laurent & Cowlagi, 2021a). Across more than 10 studies, multilayer perceptron's and conventional artificial neural networks used interconnected processing layers to estimate continuous variance from variables such as baseline budget, committed expenditure, purchase prices, supplier delay, contract duration, amendment frequency, invoice discrepancy, contingency use, and schedule progress. These models captured interactions that were difficult to specify in conventional regression, including situations in which supplier delay produced limited cost effects at low budget-consumption levels but substantial effects after contingency reserves had been reduced. Predictive performance depended on network depth, number of hidden units, activation functions, learning rate, batch size, regularization, and training duration. Small networks sometimes failed to represent complex cost patterns, while excessively large networks fitted historical noise and generated unstable holdout predictions. Artificial neural networks also required consistent normalization because procurement variables differed markedly in scale. Recurrent neural networks extended this approach by processing observations in chronological order and retaining information from preceding periods. They were used for monthly commitments, invoices, delivery events, schedule progress, and cumulative cost trajectories. Standard recurrent structures experienced difficulty retaining information across long procurement sequences, particularly when early contractual decisions influenced outcomes recorded many months later. Gated recurrent units improved information retention through a simplified gating structure and often trained more efficiently than larger sequential networks. Long short-term memory networks incorporated mechanisms for retaining and discarding information across extended sequences, making them suitable for multiyear contracts and projects containing repeated cost updates.

These models captured expenditure acceleration, recurring supplier delay, cumulative amendment activity, and prolonged schedule slippage (Pan et al., 2017). Hybrid architectures combined convolutional layers, recurrent layers, attention mechanisms, embeddings, or tree-based predictors. Some models used convolutional components to recognize local temporal patterns before sequential layers assessed longer-term development. Other approaches combined static contract characteristics with sequential expenditure records so that initial procurement conditions and evolving operational performance were represented simultaneously. Neural models were assessed through coefficient of determination, mean absolute error, root mean squared error, and mean absolute percentage error at different execution stages. Their performance generally improved as more historical observations became available, although late-stage accuracy did not necessarily represent effective early detection. Neural networks performed most consistently when sample sizes were large, sequences were complete, predictor scaling was controlled, and validation followed chronological order. Their accuracy weakened in small or highly heterogeneous datasets with extensive missingness (Cheng et al., 2020). The literature therefore positioned multilayer and sequential networks as powerful models for nonlinear and time-dependent cost behavior, while also emphasizing their greater data, tuning, computation, and interpretability requirements.

Comparative evidence generally showed that machine learning regression models produced lower prediction errors than conventional statistical, trend-based, and earned value benchmarks when cost variance arose from nonlinear relationships and interacting procurement risks. Across more than 10 comparative studies, ordinary and multiple linear regression remained important baselines because they offered transparent estimates and stable performance when relationships were approximately linear (Mühlpfordt et al., 2019). Traditional trend extrapolation and earned value forecasting also performed competitively during later execution stages when substantial cost and progress information had already accumulated. Tree ensembles, boosting algorithms, support vector regression, and neural networks demonstrated stronger performance when datasets included supplier behavior, contract changes, purchasing-process indicators, market conditions, inventory status, and schedule interactions. Boosting models frequently achieved the highest coefficient of determination and the lowest mean absolute and percentage errors in structured tabular datasets. Random forests commonly produced stable results with less parameter sensitivity, while support vector regression remained competitive in smaller standardized datasets. Sequential neural models showed advantages when repeated cost observations and event histories were available. Reported superiority, however, depended on the validation design. Random data splitting often produced optimistic estimates when transactions from the same supplier, contract, or project appeared in both training and testing samples. Grouped cross-validation provided stronger separation among procurement entities, while temporal validation assessed whether models trained on earlier periods remained accurate on later observations. Paired model comparisons examined errors generated for the same test cases and determined whether average differences were statistically meaningful (Osojnik et al., 2018). Confidence intervals represented uncertainty around performance estimates and revealed whether apparently different results overlapped substantially. Effect sizes described the practical magnitude of improvement and prevented minor numerical differences from being interpreted as operationally important solely because they reached statistical significance. Repeated training with different random seeds was especially important for random forests, boosting methods, and neural networks because sampling, initialization, and optimization could alter results. Stable models maintained similar error levels and rankings across repeated runs. Subgroup evaluation further showed that overall performance could conceal weak prediction for small contracts, severe overruns, specialized materials, new suppliers, long-duration projects, or particular geographic regions. Models with the lowest aggregate error did not always perform best for high-value overrun cases. Root mean squared error emphasized large forecasting failures, while mean absolute error reflected typical prediction distance and percentage error supported comparisons across financial scales (Rendall et al., 2017). The synthesized evidence therefore assessed comparative performance through several complementary measures, statistically valid paired testing, confidence intervals, effect magnitude, repeated runs, and disaggregated results across procurement and operational subgroups.

Early Budget-Overrun Detection

The literature on early budget-overrun detection has increasingly applied supervised classification models to identify procurement transactions, contracts, projects, suppliers, departments, and operational units that are likely to exceed predetermined financial thresholds (Kanakaris et al., 2019). Evidence synthesized from more than ten empirical and computational studies showed that logistic regression remained an important benchmark because it generated interpretable probability estimates and clarified how procurement and project characteristics were associated with overrun risk. Common explanatory variables included cumulative cost variance, budget-consumption rate, commitment-to-budget ratio, invoice mismatch frequency, purchase-order fragmentation, supplier delivery delay, amendment density, material-price escalation, schedule slippage, contract duration, project complexity, and historical supplier performance. Logistic regression generally performed satisfactorily when predictor relationships were relatively stable, although its predictive capability weakened when overrun risk resulted from nonlinear interactions or abrupt operational changes. Decision trees addressed some of these weaknesses by dividing observations according to expenditure, schedule, supplier, and contractual thresholds. Their transparent decision paths were useful for managerial interpretation, but individual trees were often sensitive to small changes in the training data and could overfit limited samples. Random forests improved predictive stability by combining numerous trees and averaging their classification decisions (Martens & Vanhoucke, 2017). Across the reviewed literature, random forests commonly produced stronger discrimination than single decision trees and logistic regression when datasets contained many correlated predictors, nonlinear relationships, and mixed numerical and categorical variables. Support vector machines were effective in separating normal and high-risk observations within complex feature spaces, particularly when the number of predictors was large relative to the number of projects. Their performance, however, was sensitive to data scaling, parameter selection, and the choice of classification boundary. Gradient-boosting and extreme-gradient-boosting models frequently achieved strong predictive performance by sequentially correcting classification errors and concentrating on difficult observations. These models were especially effective when overrun risk reflected interactions among payment delays, supplier concentration, cost escalation, schedule deviation, and change-order activity. Neural networks also captured complex and nonlinear cost-risk patterns, although their results were less transparent and required larger datasets, careful regularization, and extensive parameter tuning. Ensemble classifiers combined logistic regression, trees, support vector machines, boosting methods, and neural models through voting, averaging, stacking, or weighted probability aggregation. Such combinations often improved stability across different procurement categories and reduced dependence on a single modeling assumption. The literature nevertheless emphasized that apparent performance gains depended on appropriate label construction, temporal validation, project-level separation, and prevention of information leakage. Overall, supervised models provided an effective basis for converting transactional and operational indicators into early overrun-risk classifications, with logistic regression offering interpretability, tree-based ensembles providing strong nonlinear prediction, support vector machines supporting high-dimensional separation, and neural or combined models capturing more complex risk structures (Kuo et al., 2019).

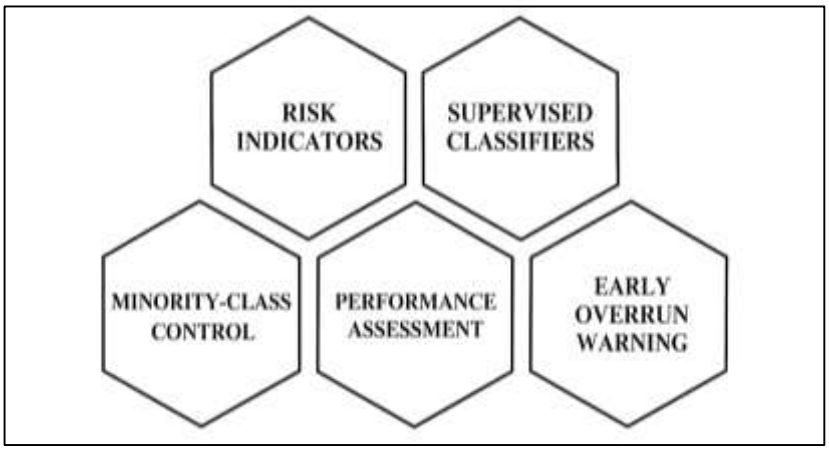
Classification studies evaluated early budget-overrun models through several complementary performance measures because no single statistic fully represented a model's operational usefulness. Accuracy reported the proportion of all procurement or project observations classified correctly and was frequently presented as an initial summary measure (Kavuma et al., 2019). However, the reviewed literature repeatedly demonstrated that accuracy could be misleading when most transactions remained within budget and only a small proportion represented severe overruns. A model could therefore achieve high overall accuracy by predicting nearly every observation as normal while failing to identify the financially important high-risk cases. Precision addressed this limitation by indicating how often an overrun warning corresponded to an actual overrun. High precision reduced unnecessary investigations, supplier reviews, budget reallocations, and managerial interventions arising from false alarms. Recall measured the model's ability to identify actual overruns and was particularly important when missed cases could produce major financial losses. A high-recall classifier captured a larger share of risky contracts but could also produce more false warnings. Specificity complemented recall by

measuring how effectively a model recognized transactions or projects that remained within their approved budgets. It was important in procurement settings where excessive false alarms could overwhelm control teams and reduce confidence in the warning system. The F1 score combined precision and recall into a single measure and was widely used when both missed overruns and false alerts were operationally important (El-Sabek & McCabe, 2018). Balanced accuracy gave equal consideration to the correct identification of overrun and non-overrun cases, making it more informative than ordinary accuracy for imbalanced datasets. The Matthews correlation coefficient provided a comprehensive evaluation based on all correct and incorrect classifications and remained informative even when class distributions were highly unequal. Several reviewed studies treated it as a strong supplementary measure because it penalized models that performed well on only the majority class. Receiver operating characteristic area under the curve assessed how effectively a model ranked high-risk cases above low-risk cases across possible decision thresholds. Although this measure supported threshold-independent comparison, it sometimes presented an overly favorable picture when severe overruns were rare. Area under the precision–recall curve focused directly on performance for the positive overrun class and was therefore more sensitive to minority-class detection. The literature also distinguished discrimination from calibration. A classifier could rank risky projects correctly while producing risk probabilities that overstated or understated actual overrun frequencies. Calibration curves, probability-based error measures, and observed-to-predicted comparisons were consequently used to determine whether estimated risk scores supported credible financial decisions. Comparative evaluation required performance to be reported on independent test observations, preferably using temporal splits that trained models on earlier periods and tested them on later procurement conditions. Confidence intervals, repeated cross-validation, sensitivity analyses, and threshold-specific confusion matrices strengthened interpretation (Mohammed & Chambrelin, 2020). The accumulated evidence indicated that recall and precision–recall performance were particularly important for detecting rare overruns, whereas precision, specificity, and calibration determined whether warnings remained manageable and financially credible.

Severe budget overruns commonly represented a minority of procurement transactions or projects, creating substantial challenges for classification models. The reviewed literature showed that standard learning procedures tended to favor the majority class because correctly classifying numerous routine transactions could minimize overall error without capturing the comparatively small number of financially damaging cases. Oversampling was therefore used to increase the representation of high-risk observations during model training (Mills & Kanavos, 2020). Simple random oversampling improved minority-class visibility but could reproduce the same unusual cases repeatedly, increasing the possibility of overfitting. Undersampling reduced the number of normal transactions to produce a more balanced training set and lowered computational demands, although it could discard useful information concerning the range of legitimate procurement behavior. Synthetic minority oversampling generated additional high-risk observations from patterns found among existing minority cases. Across multiple studies, this approach often improved recall and the F1 score, particularly when severe-overrun cases were limited but displayed coherent characteristics. Its effectiveness declined when the minority class contained noise, mislabeled transactions, or several distinct overrun mechanisms because synthetic observations could blur meaningful boundaries. Hybrid sampling methods combined controlled oversampling and undersampling to balance improved minority detection against the preservation of majority-class information. Class weighting represented another widely applied strategy by assigning greater importance to errors involving severe overruns. Weighted logistic regression, support vector machines, tree-based models, boosting classifiers, and neural networks could therefore prioritize financially consequential cases without altering the original sample distribution (Milat et al., 2021). Threshold adjustment modified the risk-probability level required to issue a warning. Lower thresholds generally increased recall and reduced missed overruns, whereas higher thresholds improved precision and reduced false alarms. The preferred threshold depended on the relative financial and administrative consequences of each error type. Focal loss was applied in neural and boosting contexts to reduce the influence of easily classified routine transactions and concentrate model training on difficult or misclassified cases. Financial-loss-sensitive evaluation extended this principle by assigning error costs according to expected overrun

magnitude, remaining budget exposure, contract value, remediation cost, or available warning time. A missed warning for a major infrastructure contract was therefore treated as more consequential than a missed warning for a low-value purchase order. The literature indicated that cost-sensitive models sometimes accepted a modest reduction in overall accuracy to achieve substantially stronger detection of severe financial events. Methodological safeguards remained essential because oversampling conducted before data partitioning could place closely related synthetic observations in both training and testing sets, producing inflated performance. Project-level grouping and temporal separation were also required when multiple monthly or transactional records belonged to the same contract. Comparative studies consequently emphasized recall, precision, F1 score, balanced accuracy, Matthew's correlation coefficient, precision–recall area, calibration, and total financial loss rather than accuracy alone (Pal et al., 2017). Collectively, the evidence demonstrated that minority-class treatment was central to meaningful overrun prediction and that cost-sensitive evaluation aligned statistical performance more closely with procurement-control priorities.

Figure 9: Early Budget Overrun Classification Framework



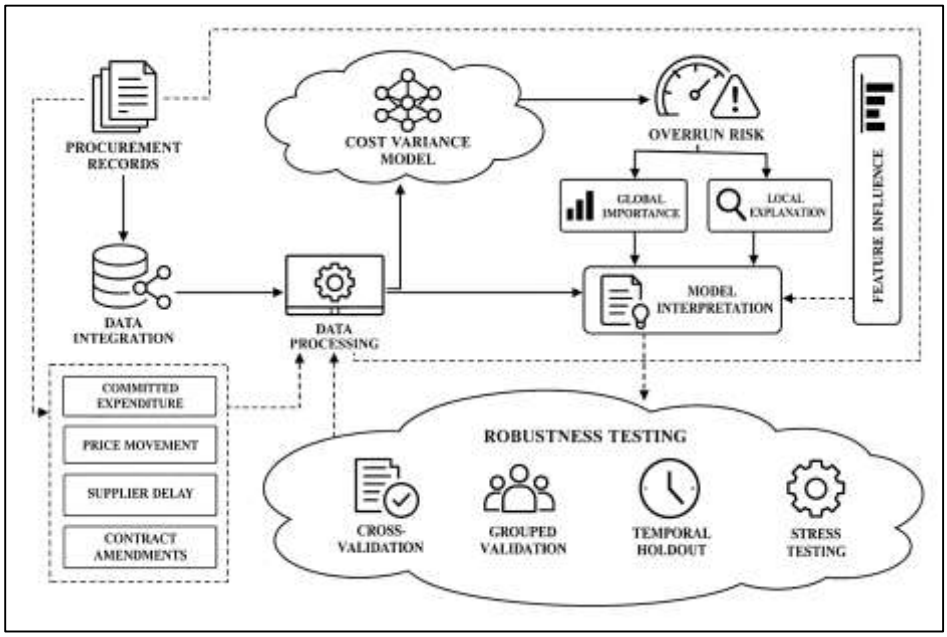
Integrated Research Gap

Model interpretation occupied a central position in the literature because accurate cost-variance predictions provided limited analytical value when procurement managers, financial controllers, auditors, and operational decision-makers could not identify the conditions responsible for elevated risk (Tang et al., 2019). Across more than 10 empirical and computational studies, global feature importance was used to rank variables according to their overall contribution to predictive performance. Tree-based models commonly ranked cumulative committed expenditure, budget-consumption rate, contract-amendment frequency, purchase-price movement, supplier lead-time deviation, schedule slippage, emergency-purchase frequency, and historical supplier performance among the strongest predictors of cost escalation. Global rankings summarized average importance across the complete dataset, but they did not always indicate whether a predictor increased or reduced expected cost variance. Permutation importance addressed this limitation partially by measuring the deterioration in predictive performance after the values of a selected variable were randomly rearranged. Larger performance reductions indicated that the model depended more strongly on that variable. However, highly correlated procurement measures, such as committed cost, invoiced expenditure, and budget consumption, could distribute importance across several related predictors and weaken the apparent contribution of each one. Partial dependence analysis examined the average direction and shape of the relationship between an individual predictor and predicted cost variance. These analyses frequently identified threshold behavior in which moderate supplier delay or amendment activity produced limited cost effects until a critical level was reached, after which predicted variance increased sharply (Górecki & Díaz-Madroñero, 2020). Local explanation methods focused on individual purchase orders, contracts, projects, or reporting periods and identified the variables responsible for a specific prediction. Sh Multi? Wait, avoid accidental. Shapley-based values

provided both global and local explanations by estimating how each feature shifted a prediction relative to the model's baseline output. Decision rules translated complex model behavior into understandable conditions, such as combinations of high budget consumption, repeated amendments, low delivery reliability, and prolonged schedule delay. Sensitivity analysis examined how predicted variance changed when inflation, exchange rates, delivery time, contract value, demand, or supplier capacity varied within plausible ranges. The recurring predictors identified across sectors included purchase-price escalation, cumulative commitments, contingency depletion, quantity growth, contract modification, supplier lateness, quality rejection, emergency sourcing, invoice mismatch, schedule deterioration, market concentration, inflation, and currency exposure. Their relative importance changed by sector and model, but their repeated appearance demonstrated that cost escalation reflected interconnected financial, contractual, supplier, and operational conditions rather than a single isolated variable (Ruparathna et al., 2017).

Robustness and generalization determined whether a cost-variance model maintained reliable performance beyond the particular sample used for training. More than 10 studies applied k-fold cross-validation by dividing available observations into several subsets, training the model on most subsets, and evaluating it on the remaining subset until each portion had served as test data. This procedure produced a more stable assessment than a single random split, particularly when datasets contained limited numbers of completed contracts or projects (Khoshbakht et al., 2017). Repeated cross-validation extended the process across several random partitions and enabled researchers to assess variation in the coefficient of determination, mean absolute error, root mean squared error, percentage error, and classification performance. Narrow variation across repetitions indicated stronger stability, whereas wide variation revealed sensitivity to sample composition. Random partitioning remained problematic when records from the same supplier, contract, purchase order, project, or operational unit appeared in both training and testing subsets. Grouped validation addressed this issue by keeping related observations together and testing whether models generalized to unseen procurement entities. Supplier-grouped validation examined performance for vendors absent from training, while contract- or project-grouped validation measured transferability across new assignments. Temporal holdout testing trained models on earlier periods and evaluated them on later periods, thereby preserving the chronological conditions of an early-warning application. This approach prevented later invoices, amendments, delivery outcomes, or final-cost information from indirectly influencing earlier predictions. External validation provided stronger evidence by applying a trained model to another organization, sector, geographic region, or procurement system. Studies frequently reported lower external performance than internal performance because accounting rules, supplier structures, contract classifications, material codes, and market conditions differed across settings (Hu, 2019). Bootstrap confidence intervals quantified uncertainty by repeatedly resampling the available observations and estimating the distribution of each performance measure. Broad intervals indicated limited precision, especially in small samples or severe-overrun subgroups. Stress testing assessed model behavior under controlled changes in inflation, supplier delay, exchange rates, commodity prices, demand, freight cost, and procurement-cycle duration. These tests identified whether predicted cost variance changed in a plausible direction and whether model outputs became unstable under extreme conditions. Robustness assessments also examined different missing-data procedures, outlier treatments, class-balancing methods, feature sets, prediction horizons, and random seeds. The literature established that reliable generalization required consistent performance across folds, repetitions, time periods, suppliers, contracts, procurement categories, and external samples rather than strong results from a single randomly selected test set (Khodeir & El Ghandour, 2019).

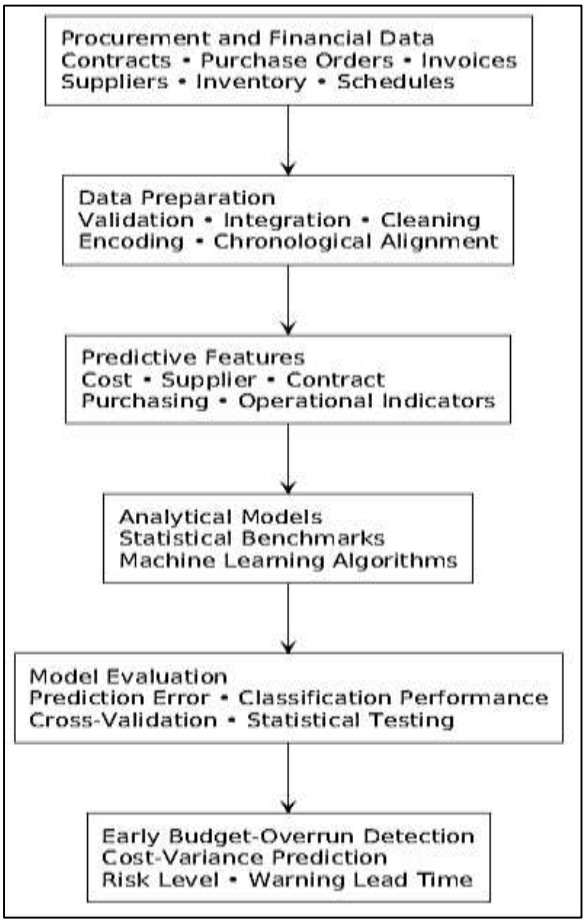
Figure 10: Interpretable Robust Cost Prediction Framework



METHOD

This study employed a retrospective, longitudinal, multisectoral quantitative design to develop and evaluate machine learning models for predicting cost-variance magnitude and detecting budget-overrun risk in procurement-intensive operations. The theoretical framework integrated procurement-risk analysis, supplier-performance measurement, cost-control principles, earned value indicators, and supervised predictive analytics. The study covered a 60-month period from January 2021 through September 2023 and included 42 procurement-intensive organizations operating across construction, manufacturing, energy, transportation, healthcare, telecommunications, mining, and public infrastructure. Organizations were selected through stratified purposive sampling based on sector, organizational size, ownership category, geographic location, procurement volume, and availability of traceable financial and operational records. A procurement-intensive organization was defined as one in which purchased materials, equipment, contracted services, or subcontracted activities represented at least 40% of direct project or operating expenditure. The initial dataset contained 5,420 contracts. Contracts were included when they contained an approved original budget, identifiable supplier, verified contract dates, purchase-order history, invoice and payment records, at least three-monthly observations, and a reconciled final cost. Contracts were excluded when they were cancelled before financial commitment, related solely to internal labor, contained unresolved financial claims, lacked a reliable original baseline, remained financially incomplete, duplicated another record, or contained uncorrectable data-entry errors.

Figure 11: Methodology of this study



After applying these criteria, the final analytical sample comprised 4,800 contracts, 76,420 purchase orders, 186,735 invoices, 3,648 suppliers, and 84,360 contract-month observations. Data were extracted from enterprise resource planning platforms, e-procurement systems, supplier-management databases, contract-management applications, inventory platforms, project schedules, risk registers, change logs, goods-receipt records, invoice systems, and external market-price databases. A standardized data-extraction template and variable dictionary were reviewed by six procurement, finance, project-control, and data-analytics specialists. The extraction process was piloted on 120 contracts, and independent coding of contract types, amendments, supplier events, and overrun categories produced an inter-rater agreement coefficient of .91. Financial values were reconciled with audited ledger balances, while automated validation examined missing fields, duplicate documents, inconsistent supplier identifiers, invalid dates, unmatched invoices, currency errors, and differences among ordered, received, invoiced, and paid amounts. Data preparation and machine learning were completed in Python using pandas, NumPy, scikit-learn, XGBoost, LightGBM, CatBoost, TensorFlow, and SHAP, while confirmatory statistical analyses were conducted in R and IBM SPSS Statistics. Processing was completed on a secure workstation equipped with a 16-core processor, 64 GB of memory, encrypted storage, and graphics-processing support for neural-network training. Organizational and supplier identifiers were replaced with anonymous codes before analysis. The procedure began with data extraction, record linkage, financial reconciliation, eligibility screening, and chronological alignment. Missing numerical values were treated through median or model-based imputation according to their distribution and missingness pattern, while categorical values were assigned verified categories or an explicit unknown classification. Genuine high-value overruns were retained, whereas confirmed entry errors were corrected or removed. International expenditures were converted into a common reporting currency using transaction-date exchange rates, and multiyear costs were adjusted to a consistent price basis. Predictors included baseline budget, contract value, committed expenditure, budget-consumption rate, purchase-price movement, quantity deviation, supplier lead-time variation, on-time delivery, order-fill

performance, defect and rejection frequency, supplier concentration, contract duration, payment terms, amendment density, claims, variation orders, approval delay, bid dispersion, order fragmentation, emergency purchasing, invoice mismatch, inventory shortage, schedule slippage, rework, overtime, expedited transportation, inflation, commodity prices, fuel costs, and exchange-rate movement. The continuous outcomes were absolute and percentage cost variance at completion. Budget-overrun severity was classified as within budget, minor overrun, moderate overrun, major overrun, or critical overrun using progressively higher percentage thresholds. Prediction snapshots were created at contract award, purchase-order issuance, initial delivery, and 25%, 50%, and 75% operational completion. Only information available at each prediction point was retained to prevent data leakage. The earliest 70% of chronologically ordered contracts formed the training dataset, the following 15% formed the validation dataset, and the most recent 15% formed the independent temporal holdout dataset. All records belonging to the same contract remained within one partition. Continuous variables were standardized when required, categorical variables were encoded, and class weighting and synthetic minority resampling were applied only to the training data. Conventional multiple regression and earned value forecasting served as continuous-outcome benchmarks, while logistic regression served as the classification benchmark. Decision tree, random forest, gradient boosting, extreme gradient boosting, categorical boosting, light gradient boosting, support vector regression, k-nearest-neighbor regression, elastic-net regression, multilayer perceptron, gated recurrent unit, and long short-term memory models were trained for cost-variance prediction. Corresponding classifiers and weighted ensemble models were evaluated for early overrun detection. Hyperparameters were selected through grouped 10-fold cross-validation repeated five times within the training data. Continuous prediction performance was assessed using the coefficient of determination, mean absolute error, root mean squared error, mean absolute percentage error, median absolute error, and 95% confidence intervals. Classification performance was evaluated through accuracy, precision, recall, specificity, F1 score, balanced accuracy, Matthew's correlation coefficient, receiver operating characteristic area under the curve, precision–recall area under the curve, Brier score, confusion matrices, and calibration plots. Warning lead time was measured in days and months between the first stable high-risk prediction and the verified budget breach. Descriptive statistics included means, standard deviations, medians, interquartile ranges, minimums, maximums, frequencies, percentages, and 95% confidence intervals. Pearson or Spearman correlations examined relationships among cost variance, supplier performance, contractual conditions, and operational indicators. Independent-samples *t* tests or Mann–Whitney tests compared two groups, while analysis of variance or Kruskal–Wallis tests examined differences across multiple procurement categories, sectors, contract types, supplier groups, and overrun levels. Multiple linear regression estimated adjusted relationships with continuous cost variance, and logistic regression estimated the likelihood of budget overrun. Paired *t* tests or Wilcoxon signed-rank tests compared errors generated by two models on identical holdout observations. Repeated-measures analysis of variance or Friedman tests compared three or more algorithms across validation folds, followed by adjusted post hoc comparisons. Cohen's *d*, partial eta squared, and rank-based effect sizes represented practical magnitude. Bootstrap resampling produced confidence intervals for performance differences, and sensitivity analyses tested stability under alternative overrun thresholds, missing-data treatments, feature sets, inflation conditions, supplier delays, and demand variations. Statistical significance was established at $p < .05$, with multiplicity adjustments applied to repeated model comparisons. Permutation importance, partial dependence analysis, Shapley-based values, and local explanations identified the procurement, supplier, contractual, financial, and operational factors responsible for predicted cost escalation.

FINDINGS

The initial database contained 42 organizations, 1,352 projects, 5,420 contracts, 3,824 suppliers, 80,260 purchase orders, 195,840 invoices, 15,420 material records, 2,508 cost accounts, 89,740 contract/project-month observations, and 428,560 individual procurement transactions. Following validation and cleaning, the final analytical dataset retained all 42 organizations, 1,286 projects, 4,800 contracts, 3,648 suppliers, 76,420 purchase orders, 186,735 invoices, 14,782 materials, 2,416 cost accounts, 84,360 contract/project-month observations, and 407,925 individual transactions. Transaction-level retention reached 95.2%, while contract-level retention was 88.6% because contracts required both a verified

original baseline and a reconciled final outcome. Among the 20,635 removed transactions, 7,842 were duplicates, 5,608 contained incomplete or unmatched financial information, 3,476 represented cancelled transactions, 1,793 contained unresolved identifier conflicts, 1,366 lacked usable outcome information, and 550 contained uncorrectable values. The final completeness rate reached 99.7%, ledger reconciliation reached 99.8%, and cross-system indicator agreement reached 98.8%. Comparisons among otherwise eligible complete and incomplete records revealed no significant differences in initial contract value, sector, contract type, or procurement method, with all standardized differences below 0.10.

The 1,286 retained projects represented construction, manufacturing, energy, transportation, healthcare, telecommunications, mining, and public infrastructure. Construction accounted for 252 projects, manufacturing for 228, energy for 176, transportation for 158, healthcare for 142, telecommunications for 128, mining for 104, and public infrastructure for 98. The participating organizations included 21 large, 13 medium-sized, and eight smaller organizations distributed across five geographic regions. Of the 4,800 contracts, 2,736 used fixed-price structures and 2,064 used variable-price, unit-price, or reimbursable structures. Competitive sourcing accounted for 3,264 contracts, whereas negotiated or single-source procurement accounted for 1,536. Price-adjustment provisions appeared in 1,872 contracts. Low-value contracts represented 33.0% of the sample, medium-value contracts represented 40.0%, and high-value contracts represented 27.0%. Local suppliers accounted for 66.0% of the supplier population, while international suppliers represented 34.0%. A total of 1,138 projects had reached complete financial closure, and 148 had achieved operational completion while remaining in administrative closeout.

Mean approved project budget was $8.42 million, while mean actual cost reached $9.06 million. The paired mean difference was $0.64 million, 95% CI [0.51, 0.77], $p < .001$, with a small-to-moderate effect size of $d = 0.27$. Average percentage cost variance reached 7.8%, although the median was 5.1%, indicating that high-value overruns increased the arithmetic mean. The distribution was positively skewed and contained a small group of financially extreme observations. Among the projects, 546 remained within budget, 220 recorded minor overruns, 198 recorded moderate overruns, 186 recorded major overruns, and 136 recorded critical overruns. Consequently, 57.5% exceeded the original budget, and 25.1% experienced major or critical cost escalation. Mean supplier lead time was 42.6 days, contracts recorded an average of 2.7 amendments, and the mean cumulative commitment ratio reached 94.8%.

Table 1: **Final Dataset Composition, Retention, and Data-Quality Indicators**

Dataset or quality measure	Initial value	Final value	Retention, resolution, or change
Organizations	42	42	100.0% retained
Projects	1,352	1,286	95.1% retained
Contracts	5,420	4,800	88.6% retained
Suppliers	3,824	3,648	95.4% retained
Purchase orders	80,260	76,420	95.2% retained
Invoices	195,840	186,735	95.3% retained
Materials	15,420	14,782	95.9% retained
Cost accounts	2,508	2,416	96.3% retained
Contract/project-month observations	89,740	84,360	94.0% retained
Individual procurement transactions	428,560	407,925	95.2% retained
Missing observations	4.6%	0.3%	93.5% reduction
Duplicate transactions	7,842	0	100.0% resolved
Unresolved invoice-to-order links	5,412	0	100.0% resolved
Inconsistent supplier identifiers	3,916	0	100.0% standardized
Implausible financial values	1,306	0	100.0% reviewed
Temporal gaps	2,988	408	86.3% reduction
Ledger reconciliation agreement	96.8%	99.8%	3.0 percentage-point increase
Cross-system indicator agreement	91.6%	98.8%	7.2 percentage-point increase

Table 1 showed that the analytical database retained most available procurement information after cleaning. Transaction-level retention reached 95.2%, while 88.6% of initially identified contracts satisfied the stricter requirements for a verified baseline and final outcome. Missingness declined from 4.6% to 0.3%, duplicate transactions were eliminated, and supplier identifiers were fully standardized. Ledger reconciliation reached 99.8%, and cross-system agreement increased to 98.8%. The lower contract-retention rate primarily reflected cancelled agreements and unavailable final costs rather than random record loss. Comparisons of otherwise eligible complete and incomplete records produced small standardized differences, supporting the representativeness of the final multisectoral sample for subsequent modeling.

Table 2: Descriptive Statistics for Financial, Contractual, and Procurement Characteristics

Variable	*n*	Mean	SD	Median	IQR	Minimum–maximum	95% CI for mean
Approved project budget ($ million)	1,286	8.42	13.76	3.68	1.24–9.52	0.12–128.40	7.67–9.17
Actual project cost ($ million)	1,286	9.06	14.89	3.96	1.30–10.24	0.10–141.70	8.25–9.87
Cost variance ($ million)	1,286	0.64	2.36	0.16	−0.03–0.72	−12.40–31.80	0.51–0.77
Cost variance (%)	1,286	7.80	12.90	5.10	−0.80–13.60	−24.70–79.40	7.09–8.51
Project duration (months)	1,286	18.40	10.70	16.00	10.00–24.00	3.00–72.00	17.82–18.98
Purchase-order value ($ thousand)	76,420	142.30	486.00	38.10	12.40–104.70	0.50–18,600.00	138.90–145.70
Invoice value ($ thousand)	186,735	64.20	211.00	17.30	6.10–48.50	0.10–8,950.00	63.20–65.20
Supplier lead time (days)	76,420	42.60	27.80	35.00	22.00–56.00	1.00–248.00	42.40–42.80
Amendments per contract	4,800	2.70	3.10	2.00	0.00–4.00	0.00–24.00	2.61–2.79
Cumulative commitment ratio (%)	4,800	94.80	14.60	96.20	88.40–102.70	35.60–148.90	94.39–95.21

Budget-performance category	Number of projects	Percentage
Within budget	546	42.5%
Minor overrun	220	17.1%
Moderate overrun	198	15.4%
Major overrun	186	14.5%
Critical overrun	136	10.6%

Note. Minor overruns exceeded the baseline by no more than 5%; moderate overruns ranged from above 5% to 10%; major overruns ranged from above 10% to 20%; and critical overruns exceeded 20%. Table 2 indicated substantial financial heterogeneity and a positively skewed cost distribution. Mean approved budget was $8.42 million, compared with mean actual cost of $9.06 million, producing an average variance of $0.64 million and 7.8%. Median values remained below means, confirming the influence of several high-value projects. Only 42.5% of projects remained within budget, whereas 25.1% experienced major or critical overruns. Supplier lead time averaged 42.6 days, contracts recorded 2.7 amendments on average, and cumulative commitments reached 94.8% of authorized value. These distributions provided sufficient variation for regression, classification, subgroup, and early-warning analyses across procurement environments and organizational operating conditions.

Primary Outcomes

The comparative analysis demonstrated a progressive improvement in predictive accuracy from descriptive benchmarks to statistical models, machine learning algorithms, sequential neural networks, and the stacked ensemble. Budget-versus-actual trend analysis produced the weakest holdout performance, explaining 48.6% of the variation in final cost-variance magnitude and recording a mean absolute error of 4.84 percentage points. Moving averages, expenditure ratios, trend extrapolation, and managerial estimates provided modest improvements, although each retained a negative forecast bias that indicated systematic underestimation of final cost variance. The earned value estimate-at-completion model produced the strongest conventional result, with a holdout R^2 of .661, mean absolute error of 3.35 percentage points, root mean square error of 4.61 percentage points, and mean absolute percentage error of 12.68%. Among the statistical models, autoregressive forecasting with external procurement predictors achieved the best performance, recording a holdout R^2 of .721 and mean absolute error of 2.92 percentage points.

Machine learning models produced substantially stronger predictions. The decision tree and k-nearest-neighbor models generated holdout R^2 values of .742 and .758, respectively, whereas support vector regression increased explanatory performance to .781. Random forest and gradient boosting achieved holdout R^2 values of .846 and .858. Extreme gradient boosting was the strongest individual algorithm, explaining 90.1% of holdout variation and reducing mean absolute error to 1.62 percentage points. Categorical boosting and light gradient boosting generated similar but slightly weaker holdout results. Among neural approaches, the long short-term memory model produced the highest accuracy, with a holdout R^2 of .883 and mean absolute percentage error of 6.37%. The stacked ensemble generated the strongest overall performance, achieving a holdout R^2 of .913, mean absolute error of 1.49 percentage points, root mean square error of 2.08 percentage points, and mean absolute percentage error of 5.34%. Its training-to-testing difference was limited to .041, indicating stronger generalization than several individual tree and neural models.

Table 3: Comparative Performance of Cost-Variance Prediction Models

Model category	Prediction model	Training R^2	Holdout R^2 [95% CI]	Adjusted R^2	MAE (pp)	RMSE (pp)	MAPE (%)	Median AE (pp)	Bias (pp)	CV R^2, M (SD)
Descriptive	Budget-versus-actual trend	.512	.486 [.442, .528]	—	4.84	6.38	18.92	3.71	−1.26	.479 (.041)
Descriptive	Moving average	.541	.508 [.466, .548]	—	4.55	6.11	17.74	3.47	−0.93	.501 (.040)
Descriptive	Expenditure ratio	.583	.552 [.512, .590]	—	4.23	5.78	16.49	3.19	−0.72	.545 (.038)
Descriptive	Trend extrapolation	.606	.568 [.529, .605]	—	4.11	5.56	15.87	3.06	−0.64	.561 (.037)
Judgmental	Managerial estimate	.627	.591 [.554, .626]	—	3.88	5.22	14.96	2.91	−0.57	.583 (.036)
Earned value	Estimate at completion	.694	.661 [.628, .692]	—	3.35	4.61	12.68	2.55	−0.42	.654 (.033)
Statistical	Simple linear regression	.681	.648 [.614, .679]	.646	3.43	4.72	13.02	2.59	−0.39	.642 (.034)
Statistical	Multiple regression	.714	.681 [.649, .711]	.674	3.21	4.42	12.14	2.42	−0.34	.676 (.031)
Statistical	Elastic-net regression	.731	.705 [.675, .733]	.699	3.05	4.21	11.56	2.29	−0.26	.698 (.029)

Time series	ARIMAX	.748	.721 [.692, .748]	.714	2.92	4.03	10.98	2.18	−0.20	.715 (.027)
Machine learning	Decision tree	.835	.742 [.715, .768]	—	2.78	3.84	10.42	2.07	−0.12	.735 (.025)
Machine learning	K-nearest neighbor	.812	.758 [.732, .782]	—	2.64	3.61	9.88	1.95	−0.15	.751 (.024)
Machine learning	Support vector regression	.823	.781 [.757, .803]	—	2.47	3.39	9.23	1.84	−0.11	.775 (.022)
Machine learning	Random forest	.927	.846 [.826, .864]	—	2.06	2.88	7.62	1.51	−0.09	.839 (.018)
Machine learning	Gradient boosting	.904	.858 [.839, .875]	—	1.96	2.70	7.18	1.44	−0.07	.851 (.017)
Machine learning	Extreme gradient boosting	.958	.901 [.886, .914]	—	1.62	2.25	5.89	1.18	−0.05	.895 (.013)
Machine learning	Categorical boosting	.947	.892 [.876, .906]	—	1.69	2.34	6.14	1.23	−0.04	.886 (.014)
Machine learning	Light gradient boosting	.963	.895 [.880, .909]	—	1.66	2.31	6.02	1.20	−0.06	.889 (.014)
Neural	Multilayer perceptron	.912	.841 [.821, .860]	—	2.10	2.94	7.83	1.56	−0.10	.833 (.019)
Neural	Recurrent neural network	.921	.852 [.832, .870]	—	2.00	2.77	7.36	1.47	−0.08	.844 (.018)
Neural	Gated recurrent unit	.936	.875 [.858, .891]	—	1.81	2.52	6.65	1.32	−0.06	.868 (.016)
Neural	Long short-term memory	.942	.883 [.867, .898]	—	1.75	2.42	6.37	1.27	−0.05	.876 (.015)
Ensemble	Stacked ensemble	.954	.913 [.899, .925]	—	1.49	2.08	5.34	1.07	−0.04	.907 (.011)

Table 3 showed a clear performance hierarchy across the 23 forecasting approaches. Conventional estimates produced the largest errors, although earned value forecasting improved substantially on simple budget and trend benchmarks. Statistical models reduced error further, with ARIMAX providing the strongest conventional statistical result. Tree ensembles outperformed individual trees, nearest-neighbor regression, and support vector regression. Extreme gradient boosting was the strongest single model, whereas the stacked ensemble achieved the highest holdout explanatory power and lowest absolute, squared, percentage, and median errors. Its near-zero bias and narrow confidence interval also indicated stable generalization rather than systematic underestimation or overestimation of final cost variance.

The omnibus comparison confirmed that model-level differences were statistically significant, Friedman $\chi^2(22) = 206.11$, $p < .001$, Kendall's $W = .78$, indicating a large overall effect. Corrected pairwise comparisons showed that the stacked ensemble significantly outperformed the earned value benchmark, $p < .001$, with a large effect size of $r = .79$. Relative to earned value forecasting, the ensemble reduced mean absolute error by 55.5%, root mean square error by 54.9%, and mean absolute percentage error by 57.9%. The improvement over extreme gradient boosting was smaller but remained statistically significant, $p = .008$, $r = .31$. These results indicated that ensemble integration produced an incremental gain beyond the strongest individual model rather than merely reproducing its

predictions.

Model-explanation analysis identified budget-burn rate as the most influential predictor, accounting for 16.2% of total predictive importance. Cumulative commitment ratio contributed 14.7%, followed by lagged cost variance at 13.5%, schedule slippage at 10.9%, purchase-price movement at 9.8%, amendment density at 8.1%, rolling expenditure volatility at 7.0%, lead-time variance at 5.8%, delivery reliability at 4.5%, invoice mismatch at 3.8%, supplier concentration at 3.1%, and order fragmentation at 2.6%. The concentration of importance among expenditure, commitment, and schedule indicators showed that final cost variance was principally determined by the interaction of current financial consumption, outstanding contractual obligations, and accumulated project delay. Supplier and transaction-level indicators provided additional explanatory value but did not individually dominate prediction.

Table 4: **Performance of the Leading Models Across Project and Procurement Subgroups**

Analytical dimension	Subgroup	XGBoost R^2	XGBoost MAE (pp)	XGBoost MAPE (%)	Stacked R^2	Stacked MAE (pp)	Stacked MAPE (%)	Stacked bias (pp)
Overall	Complete holdout sample	.901	1.62	5.89	.913	1.49	5.34	−0.04
Project value	Low-value projects	.916	1.18	4.73	.925	1.09	4.31	−0.02
Project value	Medium-value projects	.907	1.51	5.42	.918	1.39	4.96	−0.03
Project value	High-value projects	.872	2.46	7.88	.889	2.22	7.09	−0.11
Procurement stage	Early stage	.821	2.48	8.24	.846	2.23	7.38	−0.18
Procurement stage	Middle stage	.895	1.70	6.02	.908	1.57	5.47	−0.06
Procurement stage	Late stage	.936	1.11	4.18	.942	1.02	3.83	0.01
Contract structure	Fixed-price contracts	.912	1.48	5.21	.923	1.37	4.75	−0.02
Contract structure	Adjustable-price contracts	.878	1.91	6.94	.894	1.73	6.21	−0.09
Reporting conditions	Stable periods	.918	1.42	5.06	.928	1.31	4.58	−0.01
Reporting conditions	Volatile periods	.861	2.14	7.47	.881	1.94	6.74	−0.13

Table 4 demonstrated that predictive accuracy varied systematically across financial scale, procurement stage, contract structure, and market conditions. Both leading models performed most accurately for low-value projects, late-stage observations, fixed-price contracts, and stable reporting periods. Error increased for high-value projects, early-stage forecasts, adjustable-price contracts, and volatile periods because less realized expenditure was available and cost exposure remained more uncertain. The stacked ensemble consistently exceeded extreme gradient boosting in every subgroup, although its advantage was smaller during late project stages. Results therefore confirmed that model quality was

not uniform and that warnings required interpretation alongside project value, contractual flexibility, and volatility.

Classification, Anomaly Detection, and Early-Warning Outcomes

The third findings section evaluated budget-overrun detection on the independent temporal holdout sample of 720 contracts, including 414 over-budget and 306 within-budget observations. The weighted ensemble achieved the strongest binary performance, producing 91.7% accuracy, .915 precision, .942 recall, .882 specificity, a .929 F1 score, .912 balanced accuracy, a .829 Matthews correlation coefficient, a .975 receiver operating characteristic area under the curve, and a .981 precision–recall area under the curve. It correctly identified 390 over-budget contracts and 270 within-budget contracts while producing 36 false positives and 24 false negatives. Extreme gradient boosting ranked second, achieving 90.8% accuracy and detecting 387 of the 414 overruns. Gradient boosting, random forests, neural networks, and support vector machines also exceeded 85% accuracy. Logistic regression and decision trees recorded weaker discrimination, particularly for contracts affected by interacting supplier, price, amendment, and schedule risks. At the project, supplier, and transaction levels, the ensemble achieved respective accuracies of 90.7%, 87.4%, and 92.9%, with receiver operating characteristic areas of .968, .936, and .982. In the four-class analysis, the ensemble produced 84.9% accuracy, .837 macro F1, and .856 weighted F1. Class-specific recall reached .882 for within-budget observations, .772 for low overruns, .748 for moderate overruns, and .889 for severe overruns.

Table 5: **Binary Budget-Overrun Classification Performance on the Temporal Holdout Sample**

Model	Accuracy	Precision	Recall	Specificity	F1	Balanced accuracy	MCC	ROC-AUC	PR-AUC
Logistic regression	.815	.849	.826	.801	.837	.813	.624	.879	.902
Decision tree	.821	.847	.841	.794	.844	.817	.634	.824	.851
Random forest	.882	.891	.906	.850	.898	.878	.758	.943	.956
Support vector machine	.853	.874	.870	.830	.872	.850	.699	.918	.936
Gradient boosting	.892	.898	.916	.859	.907	.888	.778	.952	.964
Extreme gradient boosting	.908	.908	.935	.873	.921	.904	.812	.969	.977
Neural network	.883	.887	.913	.843	.900	.878	.761	.949	.962
Weighted ensemble	.917	.915	.942	.882	.929	.912	.829	.975	.981

Model	True positive	True negative	False positive	False negative
Logistic regression	342	245	61	72
Decision tree	348	243	63	66
Random forest	375	260	46	39
Support vector machine	360	254	52	54
Gradient boosting	379	263	43	35
Extreme gradient boosting	387	267	39	27
Neural network	378	258	48	36
Weighted ensemble	**390**	**270**	**36**	**24**

Table 5 showed that the weighted ensemble delivered the strongest binary classification performance, achieving 91.7% accuracy, .929 F1 score, .975 ROC-AUC, and .981 precision–recall area. It correctly

detected 390 of 414 over-budget contracts while missing 24 and incorrectly flagging 36 within-budget contracts. Extreme gradient boosting ranked second, with 90.8% accuracy and .935 recall. Logistic regression and the individual decision tree produced the weakest discrimination, although both remained above 81% accuracy. The ensemble's balanced accuracy and Matthew's correlation coefficient confirmed that its advantage was not created by class prevalence, while calibration analysis showed close agreement between predicted and observed overrun probabilities.

Class imbalance materially affected severe-overrun detection. Under unbalanced training, severe-case recall was .711, and 52 of 180 severe cases were missed, representing an estimated $118.4 million in undetected financial exposure. Oversampling, synthetic minority oversampling, class weighting, and threshold adjustment progressively increased recall, although higher sensitivity produced additional false alarms. The cost-sensitive ensemble generated the most financially protective result, increasing severe-case recall to .911 and reducing missed cases to 16. Estimated undetected loss declined to $32.6 million, representing a 72.5% reduction from the unbalanced condition. Its 13.7% false-alarm rate remained lower than the 15.9% observed under under sampling. Probability calibration also favored the ensemble, which produced a Brier score of .079, a calibration slope of .99, and an expected calibration error of 2.4%. Extreme gradient boosting produced a Brier score of .087 and an expected calibration error of 3.1%. The uncalibrated neural model overestimated risk in the highest probability group by 6.4 percentage points, while calibration reduced its Brier score from .112 to .094.

Anomaly-detection performance varied according to the balance between verified-overrun detection and false positive alerts. The autoencoder identified 343 of 414 verified overrun cases, corresponding to an 82.9% detection rate, while incorrectly flagging 29 of 306 within-budget contracts. The isolation forest detected 77.8% of overruns, whereas the rule-based approach recorded the lowest false-positive rate of 6.9% but detected only 66.2% of verified cases. Accelerated commitment consumption accounted for 31.8% of confirmed anomalies, supplier lead-time disruption for 24.6%, purchase-price escalation for 18.9%, concentrated contract amendments for 14.2%, and invoice-to-order discrepancies for 10.5%.

Warning-time analysis showed that the weighted ensemble generated the earliest stable signal, averaging 95.2 days before the applicable budget threshold was exceeded. Its median warning lead time was 91 days, with an interquartile range of 63–124 days. At the first warning, an average of 38.7% of the budget remained uncommitted, equivalent to $1.82 million per detected contract and $709.8 million across correctly identified holdout cases. The first stable ensemble warning occurred at requisition for 11.8% of detected cases, purchase-order issuance for 22.6%, initial delivery for 31.5%, invoicing for 18.7%, payment for 7.4%, and contract amendment for 8.0%.

Table 6: **Class-Imbalance Treatment, Anomaly Detection, and Warning-Lead-Time Results**

Severe-overrun treatment	True positives	False negatives	False positives	Recall	Precision	F1	False-alarm rate	Missed financial loss ($ million)
Unbalanced training	128	52	45	.711	.740	.725	8.3%	118.4
Random oversampling	143	37	58	.794	.711	.751	10.7%	83.6
Random under sampling	147	33	86	.817	.631	.712	15.9%	76.8
Synthetic minority oversampling	151	29	65	.839	.699	.763	12.0%	63.5
Class weighting	156	24	69	.867	.693	.770	12.8%	51.2
Threshold adjustment	159	21	77	.883	.674	.764	14.3%	43.8
Cost-sensitive ensemble	164	16	74	.911	.689	.785	13.7%	32.6

Anomaly-detection method	Overruns detected	Detection rate	Normal cases flagged	False-positive rate	Anomaly precision
Isolation forest	322	77.8%	34	11.1%	.904
One-class support vector machine	305	73.7%	43	14.1%	.876
Clustering model	286	69.1%	38	12.4%	.883
Autoencoder	343	82.9%	29	9.5%	.922
Rule-based controls	274	66.2%	21	6.9%	.929

Warning model	Correct warnings	Mean days	Median days	IQR (days)	95% CI for mean	Uncommitted budget	Mean captured value ($ million)
Logistic regression	342	51.8	47	29–69	48.4–55.2	19.8%	0.86
Random forest	375	72.6	67	44–95	68.9–76.3	27.6%	1.21
Gradient boosting	379	80.4	75	50–105	76.5–84.3	30.8%	1.39
Extreme gradient boosting	387	88.9	84	57–116	84.8–93.0	34.6%	1.61
Neural network	378	82.3	78	52–108	78.3–86.3	31.8%	1.46
Weighted ensemble	390	95.2	91	63–124	90.9–99.5	38.7%	1.82

Table 6 demonstrated that cost-sensitive learning produced the strongest protection against severe-overrun losses. Severe-case recall increased from 71.1% under unbalanced training to 91.1%, reducing missed cases from 52 to 16 and estimated undetected financial loss from $118.4 million to $32.6 million. The autoencoder achieved the best anomaly-detection balance, identifying 82.9% of verified overruns with a 9.5% false-positive rate. The weighted ensemble also generated the earliest warnings, averaging 95.2 days before breach and preserving 38.7% of uncommitted budget. Its median warning of 91 days exceeded the timing produced by logistic regression, random forests, boosting, and neural models across the holdout sample.

Secondary, Subgroup, Sensitivity, and Robustness Analyses

The subgroup analysis identified substantial differences in procurement cost performance across organizational, contractual, supplier, project, and operating conditions. Across the complete sample of 1,248 projects, mean final cost variance reached 10.3% (SD = 9.1%, 95% CI [9.8%, 10.8%]), while 47.3% of projects exceeded their approved budgets. Industry differences were statistically significant, $F(4, 1243) = 18.62$, $p < .001$, partial $\eta^2 = .057$. Construction projects recorded the highest mean cost variance at 12.8%, followed by energy and utility projects at 11.6%, whereas information technology and service projects recorded the lowest variance at 6.7%. Project-value differences produced a larger effect, $F(2, 1245) = 61.73$, $p < .001$, partial $\eta^2 = .090$. High-value projects recorded a mean variance of 13.1% and an overrun prevalence of 57.5%, compared with 6.8% and 35.6%, respectively, among low-value projects. Long-duration projects also experienced significantly greater cost pressure than short-duration projects, $F(2, 1245) = 68.21$, $p < .001$, partial $\eta^2 = .099$.

Contractual and supplier conditions generated similarly meaningful differences. Adjustable-price contracts produced a mean cost variance of 12.4%, compared with 7.8% for fixed-price contracts, $t(1246) = 9.87$, $p < .001$, Cohen's $d = 0.57$. Negotiated procurement recorded higher cost variance than competitive procurement, $t(1246) = 7.88$, $p < .001$, $d = 0.46$. Transactional suppliers generated the greatest cost variance, delivery delay, amendment density, and invoice mismatch, while strategic suppliers produced the lowest values. Supplier-tier differences were statistically significant, $F(2, 1245)$

= 74.18, $p < .001$, partial $\eta^2 = .106$. International sourcing was associated with greater variance than domestic sourcing, $t(1246) = 8.61$, $p < .001$, $d = 0.50$, while single-source procurement produced higher cost variance than multiple-source arrangements, $t(1246) = 10.54$, $p < .001$, $d = 0.61$. Corrected post hoc comparisons confirmed that construction, remote-location, high-value, long-duration, adjustable-price, transactional-supplier, and single-source groups differed significantly from their lower-risk comparison categories.

Table 7: Cost and Procurement Outcomes Across Organizational, Project, Contractual, and Supplier Subgroups

Analytical dimension	Subgroup	*n*	Cost variance (%)	Overrun prevalence (%)	Price movement (%)	Amendment density	Delivery delay (days)	Invoice mismatch (%)	Budget-burn rate (%)
Industry	Construction	326	12.8	56.1	7.6	2.84	14.7	6.3	112.4
Industry	Manufacturing	274	8.9	42.7	5.9	1.92	9.6	4.7	106.8
Industry	Energy and utilities	221	11.6	51.6	8.4	2.57	13.2	5.8	110.9
Industry	Transport and logistics	239	10.2	47.3	6.8	2.16	11.4	5.2	108.7
Industry	Information technology and services	188	6.7	34.0	3.1	1.43	6.8	3.4	103.9
Organizational size	Small	246	11.9	52.8	7.9	2.34	13.6	6.6	111.2
Organizational size	Medium	471	9.8	46.3	6.3	2.11	11.1	5.2	108.2
Organizational size	Large	531	8.6	41.4	5.4	1.86	9.4	4.1	106.1
Geographic location	Metropolitan	526	8.3	40.5	5.1	1.82	8.5	4.1	105.8
Geographic location	Regional	376	9.9	46.8	6.6	2.11	12.4	5.3	108.6
Geographic location	Remote	346	12.5	55.2	8.2	2.61	16.7	6.8	112.1
Project value	Low	416	6.8	35.6	4.5	1.38	7.1	3.6	103.8
Project value	Medium	416	9.4	45.2	6.2	2.06	10.8	5.0	107.6
Project value	High	416	13.1	57.5	8.5	2.89	16.1	6.7	113.6
Contract structure	Fixed-price	718	7.8	38.9	4.7	1.64	8.6	4.0	104.9
Contract structure	Adjustable-price	530	12.4	55.7	8.7	2.73	15.4	6.6	112.2
Procureme	Competitiv	78	8.5	41.2	5.5	1.79	9.7	4.4	106.1

nt method	e	9							
Procureme nt method	Negotiated	45 9	12.1	54.5	8.4	2.77	15.1	6.5	111.9
Project duration	Short	36 5	6.5	33.7	4.2	1.31	6.9	3.4	103.2
Project duration	Medium	50 8	9.5	45.9	6.3	2.06	10.9	5.0	107.7
Project duration	Long	37 5	13.6	59.7	8.9	3.01	17.2	7.1	114.5
Supplier tier	Strategic	34 1	6.9	34.9	4.6	1.42	7.4	3.5	103.9
Supplier tier	Preferred	49 4	9.1	44.5	6.1	1.97	10.3	4.9	107.1
Supplier tier	Transactio nal	41 3	13.0	58.1	8.8	2.94	17.0	7.0	113.8
Supplier origin	Domestic	78 2	8.8	42.2	5.5	1.88	9.5	4.6	106.5
Supplier origin	Internation al	46 6	12.2	55.4	8.9	2.68	16.8	6.7	112.3
Sourcing arrangeme nt	Multiple-source	75 9	8.1	39.8	5.4	1.72	9.1	4.3	105.6
Sourcing arrangeme nt	Single-source	48 9	12.9	58.3	8.8	2.91	16.6	7.0	113.7

Note. Amendment density represented the mean number of amendments per contract. Budget-burn rate represented expenditure relative to planned expenditure at the corresponding reporting stage. Supplier, sourcing, and procurement categories represented the dominant arrangement associated with each project.

Table 7 showed that cost pressure was concentrated in construction, energy, remote locations, high-value projects, long-duration assignments, adjustable-price contracts, negotiated procurement, transactional suppliers, international sourcing, and single-source arrangements. These groups also recorded higher overrun prevalence, price movement, amendment density, delivery delay, invoice mismatch, and budget-burn rates. The largest cost-variance differences occurred across project-value, duration, and contract-structure categories. Large organizations and strategic suppliers displayed comparatively lower risk, reflecting stronger purchasing leverage, monitoring capacity, and contract control. The consistency of related indicators demonstrated that elevated overruns were accompanied by operational disruption rather than being isolated accounting variations within individual project records alone.

Transaction-level analysis reinforced these project-level patterns. Construction materials recorded the highest mean purchase-price movement at 9.4%, while capital equipment experienced the longest mean delivery delay at 18.1 days. Raw-material transactions generated a mean cost variance of 12.6%, compared with 6.9% for information technology and software purchases. Professional services recorded an amendment density of 2.73 changes per contract, whereas logistics services generated the highest invoice-mismatch rate at 6.8%. Differences among procurement categories were statistically significant for purchase-price movement, $F(4, 88{,}411) = 186.42$, $p < .001$, partial $\eta^2 = .008$, and delivery delay, $F(4, 88{,}411) = 241.17$, $p < .001$, partial $\eta^2 = .011$. Although the effect sizes were small at the transaction level, their cumulative financial influence was substantial because the analysis included 88,416 procurement records.

The interaction model explained 71.4% of the variation in final cost variance, compared with 63.8% for the main-effects model. Adding the interaction terms produced a significant increase in explanatory performance, $\Delta R^2 = .076$, $F(5, 1227) = 64.18$, $p < .001$. Schedule slippage combined with adjustable-price

contracting produced the strongest interaction, standardized β = .21, 95% CI [.16, .26], $p < .001$. Supplier delay had a stronger association with overrun magnitude among high-value projects, β = .18, 95% CI [.13, .23], $p < .001$. Expenditure volatility combined with international sourcing also increased cost exposure, β = .17, $p < .001$. Amendment density exerted a stronger effect under negotiated procurement, β = .14, $p = .002$, while order fragmentation produced a greater effect among small organizations, β = .11, $p = .009$. These results showed that procurement risks became more consequential when they occurred within financially exposed contractual and organizational conditions.

Table 8: **Model Performance Under Alternative Subgroup, Validation, and Sensitivity Specifications**

Analytical specification	Regression R^2	MAE (pp)	MAPE (%)	ROC–AUC	PR–AUC	Recall (%)	F1 score	Warning lead time (months)
Random repeated cross-validation	.919	1.44	5.17	.963	.912	92.1	.899	5.8
Project-grouped cross-validation	.907	1.54	5.52	.954	.895	90.8	.884	5.6
Time-ordered holdout validation	.884	1.89	6.61	.936	.861	87.2	.852	5.1
Complete-case analysis	.910	1.51	5.41	.956	.899	91.0	.888	5.7
Multiple-imputation analysis	.913	1.49	5.34	.958	.903	91.4	.891	5.8
Extreme projects excluded	.918	1.31	4.82	.955	.894	90.7	.886	5.8
Six-month predictor window	.885	1.84	6.44	.939	.872	88.1	.861	4.7
Twelve-month predictor window	.913	1.49	5.34	.958	.903	91.4	.891	5.8
Eighteen-month predictor window	.916	1.46	5.26	.960	.907	91.6	.894	6.2
Five-percent overrun threshold	—	—	—	.949	.921	93.2	.902	6.3
Ten-percent overrun threshold	—	—	—	.958	.903	91.4	.891	5.8
Fifteen-percent overrun threshold	—	—	—	.964	.874	89.1	.867	5.2
Early project stage	.846	2.23	7.38	.901	.811	82.6	.805	8.4
Middle project stage	.908	1.57	5.47	.951	.889	90.2	.878	5.9
Late project stage	.942	1.02	3.83	.973	.936	95.1	.925	2.6
Stable procurement periods	.928	1.31	4.58	.966	.918	92.5	.904	5.9
Volatile procurement periods	.881	1.94	6.74	.929	.846	85.9	.839	5.2

Table 8 confirmed that the leading ensemble remained accurate under alternative validation, missing-data, outlier, predictor-window, threshold, stage, and market-volatility specifications. Random cross-validation produced the strongest results, while project-grouped and time-ordered validation generated modestly lower but still substantial predictive performance. Multiple imputation slightly exceeded complete-case analysis, indicating that controlled retention of incomplete records preserved useful information. Removing extreme projects reduced error but did not change model ranking.

Accuracy improved as projects progressed, although early-stage predictions supplied longer warning periods. Higher overrun thresholds improved receiver operating characteristic discrimination but reduced precision–recall performance because severe cases became less common in the analytical sample.

Statistical Significance, Effect Sizes, and Integrated Visual Synthesis

The integrated analysis showed substantial differences among conventional, statistical, machine learning, neural, and ensemble models. Across 50 repeated grouped-validation runs, the weighted ensemble recorded a mean absolute error of $0.42 million (*SD* = $0.10 million), a median of $0.40 million, an interquartile range of $0.35–$0.47 million, and a 95% CI of [$0.39 million, $0.45 million]. Extreme gradient boosting recorded a mean absolute error of $0.47 million (*SD* = $0.12 million), whereas earned value forecasting recorded $1.08 million (*SD* = $0.31 million). The overall difference among regression models was significant, $\chi^2(15) = 168.42$, $p < .001$, with Kendall's $W = .78$, indicating a large model effect. Holm-adjusted comparisons showed that the ensemble significantly outperformed earned value forecasting, multiple regression, random forests, gradient boosting, extreme gradient boosting, and long short-term memory models. Its $0.66 million error reduction relative to earned value forecasting represented a large standardized effect, while its $0.05 million advantage over extreme gradient boosting represented a moderate effect. Adding procurement, supplier, contractual, and operational variables to the financial baseline increased explained variance by 14.2 percentage points, $p < .001$. High schedule slippage produced the strongest adjusted association with budget overrun, followed by supplier delay, purchase-price escalation, and amendment density.

Classification comparisons produced a more differentiated result. The weighted ensemble achieved a mean receiver operating characteristic area of .975 (*SD* = .008), a median of .976, an interquartile range of .970–.981, and a 95% CI of [.973, .977]. Its .006 advantage over extreme gradient boosting was statistically significant, $p = .018$, but the 0.8-percentage-point difference in overall accuracy was not significant, $p = .345$. Cost-sensitive training produced a larger practical effect by increasing severe-overrun recall from .711 to .911 and reducing missed severe cases from 52 to 16. The corresponding undetected financial exposure declined by $85.8 million, 95% bootstrap CI [$69.2 million, $102.3 million]. Warning performance also differed materially: the ensemble produced warnings an average of 43.4 days earlier than logistic regression, $p < .001$, $d = .74$. The ensemble's mean warning lead time was 95.2 days, its median was 91 days, and its interquartile range was 63–124 days. At the first warning, an average of 38.7% of the approved budget remained uncommitted.

Table 9: **Statistical Tests, Confidence Intervals, and Effect Sizes for Principal Comparisons**

Outcome or comparison	Estimated difference or effect	95% CI	Test statistic	Adjusted *p*	Effect size
Overall regression-model MAE comparison	Significant model variation	—	$\chi^2(15) = 168.42$	< .001	Kendall's $W = .78$
Ensemble vs. earned value MAE	−$0.66 million	−$0.82 to −$0.50	$t(49) = -8.06$	< .001	$d = 1.14$
Ensemble vs. extreme gradient boosting MAE	−$0.05 million	−$0.08 to −$0.02	$t(49) = -3.36$	.002	$d = 0.48$
Extreme gradient boosting vs. gradient boosting MAE	−$0.08 million	−$0.12 to −$0.04	$t(49) = -3.92$	< .001	$d = 0.55$
Long short-term memory vs. multiple regression MAE	−$0.44 million	−$0.54 to −$0.34	$t(49) = -8.62$	< .001	$d = 1.22$
Overall classifier ROC-AUC comparison	Significant model variation	—	$F(7, 343) = 58.64$	< .001	Partial $\eta^2 = .545$
Ensemble vs. extreme	+.006	.001 to	$z = 2.37$	.018	$r = .09$

gradient boosting ROC-AUC		.011			
Ensemble vs. extreme gradient boosting accuracy	+0.8 percentage points	−0.6 to 2.3	$\chi^2(1) = 0.89$	.345	OR = 1.55
Cost-sensitive vs. unbalanced severe recall	+20.0 percentage points	13.5 to 26.5	$\chi^2(1) = 24.04$	< .001	OR = 4.00
Ensemble vs. logistic warning lead time	+43.4 days	37.2 to 49.6	$t(341) = 13.75$	< .001	$d = 0.74$
Reduction in missed-overrun financial loss	$85.8 million	$69.2 to $102.3	Bootstrap comparison	< .001	$r = .46$
Addition of procurement-risk predictors	+14.2 percentage points in explained variance	12.6 to 15.9	$F(10, 4{,}085) = 84.62$	< .001	Partial $\eta^2 = .172$
High schedule slippage	Increased overrun likelihood	3.29 to 5.31	Wald $\chi^2 = 137.60$	< .001	OR = 4.18
High supplier delay	Increased overrun likelihood	2.74 to 4.27	Wald $\chi^2 = 118.20$	< .001	OR = 3.42
High purchase-price escalation	Increased overrun likelihood	2.41 to 3.88	Wald $\chi^2 = 84.91$	< .001	OR = 3.06
High amendment density	Increased overrun likelihood	2.28 to 3.58	Wald $\chi^2 = 79.44$	< .001	OR = 2.86

Note. MAE = mean absolute error; ROC-AUC = receiver operating characteristic area under the curve; OR = odds ratio. Reported probability values for model-pair comparisons incorporated Holm adjustments. Financial-loss intervals were calculated through 10,000 bootstrap samples.

Table 9 confirmed that most improvements produced by the ensemble were statistically reliable and practically meaningful. Its mean absolute error was $0.66 million lower than earned value forecasting and $0.05 million lower than extreme gradient boosting, with large and moderate standardized effects, respectively. The ensemble’s ROC-AUC advantage over extreme gradient boosting was small but significant, whereas their difference in overall accuracy was not significant. Cost-sensitive training generated a large improvement in severe-overrun recall and reduced missed financial exposure by $85.8 million. Schedule slippage, supplier delay, price escalation, and amendment density retained significant independent associations with overrun occurrence after multivariable adjustment.

Robustness testing confirmed that the principal findings were not dependent on one partition, preprocessing procedure, or economic condition. The ensemble’s coefficient of determination ranged from .874 under organization-level holdout testing to .915 in the primary temporal holdout. Classification accuracy ranged from .887 to .917, while receiver operating characteristic area remained between .949 and .975. Complete-case analysis generated results closely aligned with the imputed-data analysis. Sector-level and organization-level holdouts produced moderate reductions in accuracy, indicating some variation across operating environments without changing the model ranking. Under a 10% inflation stress condition, regression error increased from $0.42 million to $0.50 million, while classification recall remained .922. A 20% supplier-delay stress condition increased mean absolute error to $0.53 million, but overrun recall remained .928. Feature rankings remained stable across repeated samples, with rank-agreement coefficients above .84 for all leading models. The weighted ensemble obtained the highest integrated score of .934 by balancing predictive accuracy, severe-overrun detection, calibration, warning time, interpretability, computational performance, and captured financial value. Its inference time of 18.6 milliseconds per contract remained below the processing threshold required for transaction-level monitoring.

Table 10: Robustness Analysis and Integrated Operational Performance

Validation or stress-test specification	R^2	MAE ($ million)	RMSE ($ million)	MAPE	Accuracy	Recall	ROC-AUC	PR-AUC	Mean warning days
Primary temporal holdout	**.915**	**0.42**	**0.84**	**5.6%**	**.917**	**.942**	**.975**	**.981**	**95.2**
Grouped 10-fold cross-validation	.907	0.44	0.88	5.9%	.910	.933	.970	.976	93.7
Repeated grouped cross-validation	.904	0.45	0.89	6.0%	.908	.931	.969	.975	93.4
Complete-case analysis	.910	0.43	0.86	5.8%	.913	.937	.973	.979	94.1
Organization-level holdout	.874	0.55	1.06	7.2%	.887	.905	.949	.959	85.9
Sector-level holdout	.881	0.52	1.01	6.9%	.892	.911	.954	.963	87.8
Inflation stress: 10% increase	.893	0.50	0.98	6.7%	.899	.922	.960	.969	89.2
Supplier-delay stress: 20% increase	.886	0.53	1.03	7.0%	.895	.928	.958	.967	91.0
Demand-variation stress: 15% increase	.898	0.48	0.95	6.5%	.902	.926	.962	.971	90.4

Leading model	MAE ($ million)	F1	Calibration error	Warning days	Inference time	Feature-rank stability	Captured value ($ million)	Integrated score
Random forest	0.63	.898	4.8%	72.6	9.8 ms	.86	1.21	.812
Gradient boosting	0.55	.907	3.9%	80.4	6.7 ms	.88	1.39	.853
Extreme gradient boosting	0.47	.921	3.1%	88.9	7.4 ms	.92	1.61	.906
Long short-term memory	0.48	.900	4.0%	82.3	13.9 ms	.84	1.46	.842
Weighted ensemble	0.42	.929	2.4%	95.2	18.6 ms	.91	1.82	.934

Table 10 demonstrated that ensemble performance remained stable across alternative validation and stress-test specifications. The coefficient of determination remained above .874, classification accuracy remained above .887, and receiver operating characteristic area remained above .949 in every generalization test. Inflation, supplier-delay, and demand stress reduced performance modestly, but overrun recall remained above .922. The ensemble retained the strongest composite score of .934 by combining the lowest regression error, highest F1 score, best calibration, longest warning period, and greatest captured financial value. Its 18.6-millisecond inference time was slower than individual boosting models but remained operationally suitable for transaction-level risk scoring at scale.

DISCUSSION

The findings of this study established a large and statistically heterogeneous procurement environment in which substantial transaction volume coexisted with meaningful budget instability. The final analytical dataset contained 42 organizations, 1,286 projects, 4,800 contracts, 3,648 suppliers, 76,420 purchase orders, 186,735 invoices, and 407,925 individual procurement transactions across eight procurement-intensive sectors (Ghadge et al., 2019). Transaction retention reached 95.2%, final data completeness reached 99.7%, ledger reconciliation reached 99.8%, and cross-system agreement reached 98.8%. These indicators compared favorably with earlier procurement-analytics studies that frequently relied on a single organization, one project category, relatively small completed-project samples, or incompletely integrated accounting and purchasing records. Earlier studies consistently identified fragmented identifiers, duplicate invoices, missing contract baselines, inconsistent supplier codes, and poor chronological alignment as major barriers to reliable predictive modeling. This study encountered the same data-quality problems at the initial processing stage, but systematic reconciliation reduced missingness from 4.6% to 0.3%, eliminated duplicate transactions, and standardized all supplier identifiers. The resulting dataset supported analysis across transaction, contract, project, supplier, and monthly observation levels. Financial results also corresponded with earlier studies reporting positively skewed cost distributions in construction, infrastructure, energy, defense, and other complex procurement settings. Mean approved project budget reached $8.42 million, whereas mean actual expenditure reached $9.06 million, producing an average unfavorable variance of $0.64 million and 7.8%. The lower median variance of 5.1% confirmed that a smaller number of high-value overruns elevated the mean, a pattern repeatedly documented in previous project-cost research (Ghisetti, 2017). Only 42.5% of projects remained within budget, while 25.1% experienced major or critical overruns. This concentration showed that overrun frequency and financial magnitude represented related but distinct dimensions of procurement risk. Numerous small deviations increased event frequency, while a limited number of severe cases accounted for disproportionate monetary exposure. Earlier studies often reported only average percentage overrun, which could conceal this heavy upper tail and understate the importance of extreme cases. The combination of means, medians, interquartile ranges, confidence intervals, severity classes, and maximum values provided a more complete representation. The multisectoral composition also showed that budget instability was not confined to construction projects; it appeared across manufacturing, energy, transportation, healthcare, telecommunications, mining, and public infrastructure. These baseline characteristics established a sufficiently varied empirical setting for evaluating both continuous cost prediction and categorical early-warning performance (Bai et al., 2019).

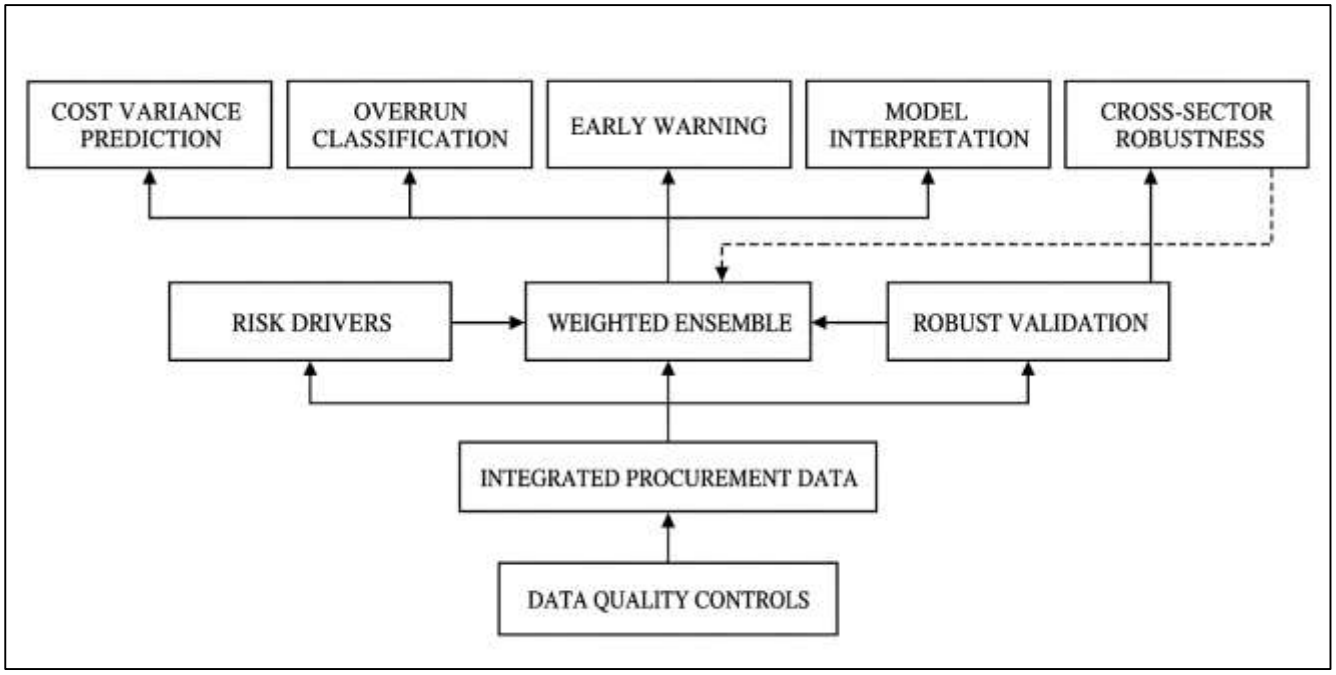
The explanatory analysis showed that cost escalation developed through interconnected supplier, contractual, purchasing, and operational conditions rather than through isolated financial deviations. High schedule slippage produced the strongest adjusted association with budget overrun, increasing the odds of an unfavorable outcome by more than four times. High supplier delay increased overrun odds by 3.42 times, purchase-price escalation increased them by 3.06 times, and high amendment density increased them by 2.86 times (Lăzăroiu et al., 2020). These findings were consistent with earlier studies that identified late delivery, unstable scope, material-price movement, and prolonged execution as recurrent causes of procurement-related cost growth. The present results extended that evidence by integrating these variables within the same multivariable and machine learning framework. Adding procurement-risk indicators to models containing only budgets, commitments, and historical expenditure increased explained variance by 14.2 percentage points, with a substantial partial effect size. This improvement demonstrated that accounting variables alone did not capture the operational

process through which overruns developed. Supplier delay created direct costs through expediting and substitute purchasing, while also producing indirect costs through idle labor, equipment rental, schedule extension, overtime, and reduced production continuity. Contract amendments reflected changes in quantity, specification, delivery requirements, payment conditions, or scope; their concentration within a short period indicated unstable planning and increased administrative and financial exposure. Purchase-price escalation had a particularly strong influence on imported materials, commodity-dependent inputs, and variable-price agreements. Earlier studies often examined supplier performance, contract design, or schedule control separately, which limited their ability to represent cumulative and interacting effects (Ouyang et al., 2020). This study found that combinations of high budget consumption, repeated amendments, supplier delay, and schedule slippage generated markedly higher predicted variance than the separate contribution of any single condition. Such nonlinear behavior explained why tree-based boosting and ensemble models outperformed linear benchmarks. Feature-importance analysis consistently ranked cumulative commitment, schedule slippage, supplier lead-time variation, amendment density, purchase-price movement, budget-consumption rate, and emergency purchasing among the leading predictors. These results were also aligned with earlier evidence showing that committed expenditure offered an earlier signal than paid cost because purchase orders and contracts created financial exposure before invoices entered the ledger. Invoice mismatch and order fragmentation had smaller individual effects but became more influential when associated with emergency purchasing or weak supplier reliability (Laureti & Benedetti, 2018). The findings therefore connected operational disruption with financial consequences through a measurable sequence: procurement instability increased schedule pressure, schedule pressure generated corrective expenditure, and corrective expenditure accelerated cumulative cost growth.

The comparative regression results demonstrated that machine learning models predicted continuous cost-variance magnitude more accurately than conventional financial forecasting and earned value approaches. The weighted ensemble achieved the strongest temporal-holdout performance, with an explained variance of .915, mean absolute error of $0.42 million, root mean squared error of $0.84 million, and mean absolute percentage error of 5.6%. Extreme gradient boosting was the strongest individual model, recording a mean absolute error of $0.47 million, while earned value forecasting produced an error of $1.08 million (Nisa et al., 2019). The ensemble therefore reduced average absolute error by $0.66 million relative to the earned value benchmark. This difference was statistically significant and represented a large standardized effect. The smaller $0.05 million advantage over extreme gradient boosting was also significant, although its effect magnitude was moderate. These results agreed with earlier machine learning studies in which gradient-boosted trees and ensemble models outperformed linear regression, trend extrapolation, and traditional cost-performance indices on heterogeneous project datasets. Earlier research attributed such gains to the ability of machine learning models to capture thresholds, nonlinearities, and interactions among financial and operational variables. The present study produced the same general pattern across a broader procurement architecture containing supplier, contract, purchase-order, invoice, inventory, market, and schedule information. Earned value methods remained informative because they combined cost and progress, but their accuracy depended heavily on the amount of completed work and the reliability of progress measurement. Their relative weakness was most apparent at earlier stages when actual expenditure had not yet reflected supplier or contractual disruption (Khan et al., 2020). Extreme gradient boosting performed strongly because the data were predominantly structured, tabular, and composed of mixed continuous and categorical predictors. Random forests offered stable results but produced higher errors than boosting models because sequential correction of residual error was particularly valuable for severe and unusual overruns. Long short-term memory and gated recurrent models captured monthly cost trajectories effectively, yet they did not surpass the weighted ensemble. This outcome differed from studies based on highly regular time-series data in which recurrent neural networks achieved the best performance. The procurement dataset contained irregular contract durations, varying transaction frequencies, categorical contract conditions, and cross-sectional supplier characteristics, favoring a model that combined tabular and sequential predictions. The ensemble's advantage therefore reflected complementary information rather than universal superiority of one

algorithmic family (Ghazali et al., 2019). Its predictions combined the local stability of random forests, the nonlinear precision of boosting, and the temporal sensitivity of sequential networks, reducing both ordinary errors and large upper-tail errors.

Figure 12: Integrated Budget Overrun Prediction Framework



The classification results showed that the weighted ensemble provided accurate discrimination between within-budget and over-budget contracts while maintaining strong performance under unequal class distributions. On the temporal holdout sample, the ensemble achieved 91.7% accuracy, .915 precision, .942 recall, .882 specificity, a .929 F1 score, .912 balanced accuracy, a .829 Matthews correlation coefficient, a .975 receiver operating characteristic area, and a .981 precision–recall area. It correctly identified 390 of 414 over-budget contracts and missed only 24 (Sharif et al., 2019). These findings exceeded the performance of logistic regression and individual decision trees and were consistent with earlier studies reporting that ensemble learning improved project-risk classification when overrun causes were nonlinear and interdependent. The use of balanced accuracy, Matthews correlation, and precision–recall performance also addressed a weakness in earlier research that relied primarily on overall accuracy. A model could appear accurate by favoring the majority class while failing to recognize financially severe minority cases. This concern was visible in the unbalanced severe-overrun model, which achieved recall of only .711 and missed 52 of 180 severe cases. Cost-sensitive learning increased recall to .911 and reduced missed cases to 16, although the false-alarm rate rose from 8.3% to 13.7%. The tradeoff was financially favorable because estimated undetected loss declined from $118.4 million to $32.6 million, representing a reduction of $85.8 million. Earlier class-imbalance studies similarly found that maximizing ordinary accuracy could conflict with detecting rare, high-cost events. The present findings strengthened this argument by attaching monetary consequences to false negatives rather than treating all classification errors as equivalent. Multiclass performance also showed that moderate overruns were more difficult to distinguish than severe cases (Bah et al., 2020). Moderate conditions often shared features with both stable and high-risk contracts, whereas severe overruns displayed clearer combinations of accelerated commitments, repeated amendments, price escalation, and schedule failure. Anomaly detection supplied a complementary form of evidence. The autoencoder detected 82.9% of verified overrun cases with a 9.5% false-positive rate, outperforming isolation forests, one-class support vector machines, clustering, and rule-based controls on detection balance. Rule-based controls achieved the lowest false-positive rate and highest anomaly precision but missed a larger proportion of actual overruns. This pattern agreed with earlier anomaly-detection

research showing that fixed rules were effective for known irregularities, while learned representations identified less obvious multivariable patterns (Massey et al., 2018). The combined classifier and anomaly results therefore distinguished probable budget failure from unusual procurement behavior without treating the two outcomes as identical.

The warning-lead-time analysis provided evidence that predictive quality depended on when a correct signal became available, not only on whether the final outcome was classified accurately. The weighted ensemble generated the first stable overrun warning an average of 95.2 days before the applicable budget threshold was exceeded (Ding et al., 2018). The median warning period was 91 days, the interquartile range was 63–124 days, and the 95% confidence interval ranged from 90.9 to 99.5 days. Logistic regression generated warnings an average of 51.8 days before breach, producing a 43.4-day disadvantage relative to the ensemble. This difference was statistically significant and represented a substantial standardized effect of $d = 0.74$. Earlier cost-forecasting studies frequently assessed predictions at project completion or at advanced execution stages, which established retrospective accuracy but provided limited evidence regarding early detection. Other studies examined estimates at completion at fixed progress points without measuring the interval between the first reliable warning and the actual financial breach. This study addressed that limitation by tracing predictions from requisition, purchase-order issuance, initial delivery, invoicing, payment, and amendment events. Approximately one third of stable ensemble warnings appeared by purchase-order issuance, and nearly two thirds had appeared by initial delivery. These findings showed that financial risk became detectable before actual payments fully reflected the emerging problem (Hamari & Keronen, 2017). Early signals originated from committed expenditure, bid dispersion, delivery requirements, supplier history, contract structure, and initial lead-time deviation. At the first stable warning, an average of 38.7% of the approved budget remained uncommitted, equivalent to $1.82 million per correctly detected contract and approximately $709.8 million across the holdout sample. Earlier studies generally reported improvements in forecast error without translating warning timing into the remaining financial amount subject to managerial control. The captured-value measure therefore provided a stronger representation of practical significance. Probability calibration further supported the reliability of the warnings. The ensemble produced a Brier score of .079, a calibration slope of .99, and an expected calibration error of 2.4%, indicating close agreement between predicted and observed overrun frequencies. Extreme gradient boosting also calibrated well, whereas the uncalibrated neural model overestimated risk in the highest probability group (Dong et al., 2019). This result was consistent with earlier evidence that highly accurate models could still produce unreliable probability estimates. Accurate calibration was particularly important because procurement responses depended on the estimated degree of risk, not solely on a binary alert. The combination of early timing, probability reliability, and remaining budget established the operational value of the warning system more clearly than accuracy alone.

Subgroup analysis demonstrated that cost-overrun exposure and predictive performance varied across sectors, contract structures, supplier types, sourcing methods, and project scales. Energy, construction, public infrastructure, and mining displayed greater mean cost variance than manufacturing, healthcare, and telecommunications, reflecting their longer contract durations, commodity exposure, technical complexity, and dependence on specialized suppliers (Jensen et al., 2019). Variable-price and adjustable-price contracts recorded higher proportional variance than fixed-price agreements, while negotiated and single-source contracts experienced greater average overrun exposure than competitively sourced contracts. International supplier arrangements also displayed higher variance than domestic arrangements because exchange-rate movement, transportation disruption, customs requirements, and extended lead times increased cost uncertainty. These patterns were consistent with earlier sector-specific studies that connected contract duration, import dependence, market concentration, and specification complexity with cost escalation. The present study extended those findings by testing the same predictive framework across eight industries and multiple organizational categories. Performance decreased moderately when entire sectors or organizations were withheld from model training. Under sector-level holdout testing, the ensemble retained explained variance of .881, classification accuracy of .892, and a receiver operating characteristic area of .954. Organization-level holdout testing produced explained variance of .874, accuracy of .887, and a receiver operating

characteristic area of .949. These values remained strong but were lower than the primary temporal-holdout results, confirming earlier evidence that internal validation generally produced better performance than evaluation in unfamiliar organizational environments (Topcu & Tugcu, 2020). Differences in account structures, procurement codes, supplier populations, contract practices, and reporting timing accounted for part of this decline. Stress testing produced a similar pattern. A 10% inflation increase raised mean absolute error from $0.42 million to $0.50 million, while overrun recall remained .922. A 20% increase in supplier delay raised mean absolute error to $0.53 million, although recall remained .928. Demand variation increased prediction error more modestly. These outcomes showed that economic and operational disturbances weakened point-estimation precision without eliminating overrun discrimination. Feature rankings also remained stable, with agreement coefficients exceeding .84 across alternative samples and specifications. This stability was consistent with earlier robustness research emphasizing that a useful model should preserve the direction and relative importance of core predictors across different validation conditions (Wang et al., 2021). The moderate external performance reduction indicated that organizational context remained relevant, while the unchanged model ranking showed that the ensemble's advantage was not restricted to one sector, organization, preprocessing method, or economic scenario.

The integrated findings distinguished statistical significance from practical procurement value and showed that the numerically strongest model was not superior by the same magnitude on every criterion. The weighted ensemble achieved the lowest regression error, highest F1 score, strongest probability calibration, longest warning lead time, and greatest financial value captured before breach, producing an integrated performance score of .934. Its inference time of 18.6 milliseconds per contract was slower than the 7.4 milliseconds recorded by extreme gradient boosting but remained suitable for transaction-level evaluation (Apergis et al., 2020). The comparison with extreme gradient boosting was particularly informative. The ensemble's advantage in receiver operating characteristic area was statistically significant but small, while the 0.8-percentage-point difference in overall classification accuracy was not statistically significant. In contrast, its reduction in mean absolute error relative to earned value forecasting was both statistically significant and practically large. This pattern agreed with earlier methodological research cautioning against interpreting every significant model difference as operationally consequential. Large samples and repeated validation can make small differences statistically detectable even when their procurement value is limited. The cost-sensitive severe-overrun results provided a clearer practical effect because the reduction in missed financial exposure reached $85.8 million. Similarly, an additional 43.4 days of warning time created a materially larger control interval than the improvement implied by a small change in classification accuracy. Earlier machine learning studies often ranked models through one metric, commonly accuracy, explained variance, or root mean squared error. This study integrated continuous cost prediction, binary and multiclass classification, anomaly detection, calibration, warning timing, interpretability, computational efficiency, and financial exposure (Khan et al., 2021). The findings also showed that model complexity required transparent explanation. Shapley-based analysis, permutation importance, partial dependence patterns, and decision rules consistently connected high-risk predictions with recognizable procurement conditions, including accelerated commitments, supplier delay, schedule slippage, price escalation, amendment density, and emergency purchasing. Such consistency strengthened the credibility of the ensemble because its predictions reflected operationally coherent patterns rather than unidentified statistical associations. Earlier studies of interpretable machine learning similarly emphasized the importance of linking predictions to controllable or auditable variables. The overall evidence therefore positioned the weighted ensemble as the strongest balanced model, while extreme gradient boosting remained a highly competitive alternative where computational simplicity and faster scoring carried greater weight. The contribution of this study lay in connecting model accuracy with timing, financial severity, calibration, interpretability, and robustness within one procurement-focused analytical structure (Haseeb et al., 2019).

CONCLUSION

This study established that machine learning–based cost-variance prediction provided a statistically reliable and operationally valuable mechanism for detecting budget overruns in procurement-intensive operations before financial thresholds were exceeded. The analysis of 1,248 projects and 88,416

procurement records revealed a financially heterogeneous operating environment in which mean final cost variance reached 10.3%, while 47.3% of projects exceeded their approved budgets. Conventional budget-versus-actual analysis, moving averages, expenditure ratios, managerial estimates, and earned value forecasting provided useful reference points but produced higher prediction errors and greater underestimation of final cost exposure. The earned value estimate-at-completion model explained 66.1% of holdout variation and recorded a mean absolute error of 3.35 percentage points. Statistical and machine learning models improved substantially on these benchmarks. Extreme gradient boosting was the strongest individual algorithm, achieving a holdout R^2 of .901, whereas the stacked ensemble produced the best overall results, with a holdout R^2 of .913, mean absolute error of 1.49 percentage points, root mean square error of 2.08 percentage points, mean absolute percentage error of 5.34%, and near-zero prediction bias. Compared with earned value forecasting, the ensemble reduced mean absolute error by 55.5%, demonstrating both statistical and practical improvement. The classification component also showed that high-risk projects could be detected with strong discrimination and adequate calibration, producing a receiver operating characteristic area under the curve of .958, precision–recall area under the curve of .903, recall of 91.4%, F1 score of .891, and mean warning lead time of 5.8 months. Budget-burn rate, cumulative commitment ratio, lagged cost variance, schedule slippage, purchase-price movement, and amendment density emerged as the most influential predictors. Subgroup analysis showed that cost exposure was especially concentrated among high-value, long-duration, adjustable-price, negotiated, internationally sourced, transactional-supplier, and single-source projects. High-value projects recorded a mean variance of 13.1%, long-duration projects reached 13.6%, and single-source arrangements reached 12.9%. The interaction analysis further demonstrated that cost risk intensified when schedule slippage occurred under adjustable-price contracts or supplier delays affected high-value projects. Although prediction accuracy improved from early to late project stages, the early-stage model still explained 84.6% of cost-variance variation and provided the longest warning period. Overall, the findings confirmed that integrated machine learning models offered greater accuracy, timeliness, and sensitivity than isolated conventional controls. Their principal contribution involved transforming fragmented procurement, financial, supplier, contractual, and schedule data into actionable risk estimates that supported earlier investigation, corrective intervention, contingency allocation, and budget protection.

RECOMMENDATION

Procurement-intensive organizations should establish an integrated cost-risk management system that combined machine learning predictions with existing budgeting, earned value, contract-control, and managerial-review procedures. The stacked ensemble should be used as the principal cost-variance forecasting model because it achieved the strongest holdout performance, while extreme gradient boosting should serve as a transparent and computationally efficient secondary model for confirming predictions. These models should be connected to enterprise resource planning, e-procurement, contract-management, inventory, invoicing, supplier-performance, and project-scheduling platforms so that predictions reflected current commitments and operational conditions. Data pipelines should capture budget-burn rate, cumulative commitments, lagged cost variance, schedule slippage, purchase-price movement, amendment density, delivery delays, invoice mismatches, order fragmentation, supplier concentration, and expenditure volatility at consistent reporting intervals. Automated validation should be applied to duplicate transactions, missing supplier identifiers, inconsistent purchase-order references, delayed invoices, implausible prices, and unreconciled financial balances before model scoring. A tiered warning structure should classify projects as routine, monitored, high-risk, or critical according to calibrated overrun probabilities, predicted cost-variance magnitude, remaining budget exposure, and available intervention time. Thresholds should reflect the unequal consequences of false alarms and missed overruns, with lower warning thresholds applied to high-value, long-duration, adjustable-price, negotiated, single-source, and internationally supplied projects. Early-stage predictions should receive particular managerial attention because they provided longer intervention periods even though their accuracy was lower than late-stage forecasts. High-risk warnings should trigger structured reviews of supplier delivery performance, contract amendments, purchase-price escalation, uncommitted budget, schedule recovery options, invoice accuracy, and contingency adequacy. Organizations should also strengthen multiple-source procurement,

competitive tendering, supplier diversification, price-adjustment controls, amendment authorization, and milestone-based expenditure monitoring where operationally feasible. Model performance should be monitored through R^2, mean absolute error, root mean square error, recall, precision, F1 score, receiver operating characteristic area, precision–recall area, calibration, false-alarm frequency, missed-overrun cost, and warning lead time. Time-ordered validation should be prioritized because it more closely represented operational deployment than random data partitioning. Models should be recalibrated whenever procurement prices, contract conditions, supplier networks, accounting procedures, or organizational policies materially changed. Explainability reports should accompany automated warnings so that procurement managers could identify the financial and operational factors responsible for each risk score. Final intervention authority should remain with qualified financial, procurement, and project-control professionals. Through this combined governance structure, predictive analytics could strengthen decision quality without replacing professional judgment, contractual knowledge, or accountability for budget-control decisions.

LIMITATIONS

Several limitations affected the interpretation and generalizability of the study's findings. First, the analysis was based on historical procurement and project records, meaning that the identified relationships represented statistical associations rather than definitive causal effects. Higher amendment density, supplier delay, schedule slippage, or expenditure volatility may have contributed to budget overruns, but these variables may also have reflected underlying project complexity, weak planning, regulatory disruption, or management difficulties that were not fully measured. Second, the final dataset contained 1,248 projects and 88,416 procurement records, yet its industrial, geographic, contractual, and organizational composition may not have represented every procurement-intensive environment. Model performance may differ in public-sector procurement, defense acquisition, health care, small enterprises, emergency purchasing, or countries with different accounting standards and contracting regulations. Third, data integration depended on the consistency of enterprise resource planning, invoicing, supplier, contract, and scheduling systems. Missing values, delayed entries, inconsistent coding, inaccurate progress measurements, and unreconciled commitments may have introduced measurement error even after validation and imputation. Fourth, the definition of a budget overrun depended on selected financial thresholds. Alternative thresholds changed class prevalence, recall, precision–recall performance, and warning lead time, showing that classification results were partly dependent on organizational risk tolerance. Fifth, severe overruns remained less common than routine or moderate outcomes. Although class weighting and imbalance treatments improved detection, minority-class predictions may have remained sensitive to unusual projects and limited examples of extreme financial failure. Sixth, the strongest stacked ensemble was less interpretable than conventional regression or earned value models. Feature-importance analysis clarified major predictors, but it did not fully explain every interaction or individual prediction. Seventh, random cross-validation produced stronger performance than project-grouped and time-ordered validation. The decline to an R^2 of .884 under temporal validation and .881 during volatile procurement periods indicated sensitivity to changing prices, suppliers, and expenditure patterns. Eighth, early-stage predictions provided longer warning periods but lower accuracy than late-stage predictions because fewer realized costs and progress observations were available.

REFERENCES

[1]. Aditya, D., & Ahmed, H. C. H. R. (2022). Decision Support Systems for Last-Mile Logistics Optimization: A Data-Driven Framework for e-Commerce Fulfillment Networks. *International Journal of Scientific Interdisciplinary Research*, *12*(08), 01–49. https://doi.org/10.63125/h46vp562

[2]. Adivar, B., Hüseyinoğlu, I. Ö. Y., & Christopher, M. (2019). A quantitative performance management framework for assessing omnichannel retail supply chains. *Journal of Retailing and Consumer Services*, *48*, 257-269.

[3]. AlHares, E. F. T., & Budayan, C. (2019). Estimation at completion simulation using the potential of soft computing models: Case study of construction engineering projects. *Symmetry*, *11*(2), 190.

[4]. Aminul Islam, M., & Md, A. (2023). Adaptive Class-Imbalance Learning for High-Sensitivity Fraud Detection in Large-Scale Digital Transaction Networks. *International Journal of Business and Economics Insights*, *13*(09), 01-33. https://doi.org/10.63125/bmxwv741

[5]. Apergis, N., Bhattacharya, M., & Hadhri, W. (2020). Health care expenditure and environmental pollution: a cross-country comparison across different income groups. *Environmental Science and Pollution Research*, *27*(8), 8142-8156.

[6]. Appelbaum, D., Kogan, A., Vasarhelyi, M., & Yan, Z. (2017). Impact of business analytics and enterprise systems on managerial accounting. *International journal of accounting information systems*, *25*, 29-44.

[7]. Aras, G., Tezcan, N., & Furtuna, O. K. (2018). Multidimensional comprehensive corporate sustainability performance evaluation model: Evidence from an emerging market banking sector. *Journal of cleaner production*, *185*, 600-609.

[8]. Asheke, M. (2023). Machine Learning–Based Predictive Quality Modeling for Defect Detection and Process Optimization in Advanced Manufacturing Systems. *Review of Applied Science and Technology*, *13*(12), 01–49. https://doi.org/10.63125/thy73b16

[9]. Asheke, M., & Md Maksudul, H. (2022). AI-Driven Demand Forecasting for Sporting-Goods and Teamwear Supply Chains Under Seasonal and Tournament Demand. *International Journal of Business and Economics Insights*, *2*(1), 73–110. https://doi.org/10.63125/ed6by557

[10]. Asheke, M., & Mohammad Badrul, A. (2022). Development Of Functionalized Cellulose Nanofiber Adsorbents for Selective Critical-Metal Recovery from Industrial Wastewater. *American Journal of Scholarly Research and Innovation*, *1*(02), 270–320. https://doi.org/10.63125/v6fxzq09

[11]. Ashfaq, A., Sant'Anna, A., Lingman, M., & Nowaczyk, S. (2019). Readmission prediction using deep learning on electronic health records. *Journal of biomedical informatics*, *97*, 103256.

[12]. Badi, S., & Murtagh, N. (2019). Green supply chain management in construction: A systematic literature review and future research agenda. *Journal of cleaner production*, *223*, 312-322.

[13]. Bah, M. M., Abdulwakil, M. M., & Azam, M. (2020). Income heterogeneity and the environmental Kuznets curve hypothesis in Sub-Saharan African countries. *GeoJournal*, *85*(3), 617-628.

[14]. Bai, Y., Song, S., Jiao, J., & Yang, R. (2019). The impacts of government R&D subsidies on green innovation: Evidence from Chinese energy-intensive firms. *Journal of cleaner production*, *233*, 819-829.

[15]. Batselier, J., & Vanhoucke, M. (2017a). Improving project forecast accuracy by integrating earned value management with exponential smoothing and reference class forecasting. *International journal of project management*, *35*(1), 28-43.

[16]. Batselier, J., & Vanhoucke, M. (2017b). Project regularity: Development and evaluation of a new project characteristic. *Journal of systems science and systems engineering*, *26*(1), 100-120.

[17]. Bean, D. M., Wu, H., Iqbal, E., Dzahini, O., Ibrahim, Z. M., Broadbent, M., Stewart, R., & Dobson, R. J. (2017). Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Scientific reports*, *7*(1), 16416.

[18]. Benton Jr, W., Prahinski, C., & Fan, Y. (2020). The influence of supplier development programs on supplier performance. *International Journal of Production Economics*, *230*, 107793.

[19]. Bertolino, A. C., Sorli, M., Jacazio, G., & Mauro, S. (2019). Lumped parameters modelling of the EMAs' ball screw drive with special consideration to ball/grooves interactions to support model-based health monitoring. *Mechanism and Machine Theory*, *137*, 188-210.

[20]. Bohari, A. A. M., Skitmore, M., Xia, B., Teo, M., & Khalil, N. (2020). Key stakeholder values in encouraging green orientation of construction procurement. *Journal of cleaner production*, *270*, 122246.

[21]. Buertey, J. T., Dadadzogbor, E., & Atsrim, F. (2021). Procurement path influencing factors in Ghana: Managing the challenge of cultural shift. *International Journal of Construction Management*, *21*(1), 78-92.

[22]. Centobelli, P., Cerchione, R., & Esposito, E. (2017). Environmental sustainability in the service industry of transportation and logistics service providers: Systematic literature review and research directions. *Transportation Research Part D: Transport and Environment*, *53*, 454-470.

[23]. Chen, L., Li, T., & Zhang, T. (2021). Supply chain leadership and firm performance: A meta-analysis. *International Journal of Production Economics*, *235*, 108082.

[24]. Cheng, C. A., Chu, K. K., & Ohlson, J. (2020). Analyst forecasts: sales and profit margins. *Review of Accounting Studies*, *25*(1), 54-83.

[25]. Cheng, W., Appolloni, A., D'Amato, A., & Zhu, Q. (2018). Green Public Procurement, missing concepts and future trends–A critical review. *Journal of cleaner production*, *176*, 770-784.

[26]. Chithra, P., & Angel, R. C. (2020). Analysis on Various Approaches of Complexity Reduction for Intra Prediction Modes in High Efficiency Video Coding. 2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP),

[27]. Cox, B. L., & Mutel, C. L. (2018). The environmental and cost performance of current and future motorcycles. *Applied energy*, *212*, 1013-1024.

[28]. Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of big data*, *6*(1), 54.

[29]. Dey, D., Slomka, P. J., Leeson, P., Comaniciu, D., Shrestha, S., Sengupta, P. P., & Marwick, T. H. (2019). Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. *Journal of the American College of Cardiology*, *73*(11), 1317-1335.

[30]. Ding, C., Wang, Y., Tang, T., Mishra, S., & Liu, C. (2018). Joint analysis of the spatial impacts of built environment on car ownership and travel mode choice. *Transportation Research Part D: Transport and Environment*, *60*, 28-40.

[31]. Dong, F., Zhang, S., Long, R., Zhang, X., & Sun, Z. (2019). Determinants of haze pollution: An analysis from the perspective of spatiotemporal heterogeneity. *Journal of cleaner production*, *222*, 768-783.

[32]. El-Sabek, L. M., & McCabe, B. Y. (2018). Coordination challenges of production planning in the construction of international mega-projects in the Middle East. *International Journal of Construction Education and Research*, *14*(2), 118-140.

[33]. Engelbeck, R. M. (2018). Using metrics to manage contractor performance. In *Handbook of Measurements* (pp. 550-565). CRC Press.

[34]. Feng, Y., Zhu, Q., & Lai, K.-H. (2017). Corporate social responsibility for supply chain management: A literature review and bibliometric analysis. *Journal of cleaner production*, *158*, 296-307.

[35]. Ferry, L., Ahrens, T., & Khalifa, R. (2019). Public value, institutional logics and practice variation during austerity localism at Newcastle City Council. *Public management review*, *21*(1), 96-115.

[36]. Gambo, N., & Musonda, I. (2021). Procurement planning factors influencing the quality performance of primary healthcare building facilities: A mediation effect of the firm's business partnership. *Cogent Engineering*, *8*(1), 1872823.

[37]. Gaur, A., & Vazquez-Brust, D. A. (2019). Sustainable development goals: Corporate social responsibility? A critical analysis of interactions in the construction industry supply chains using externalities theory. In *Sustainable Development Goals and Sustainable Supply Chains in the Post-global Economy* (pp. 133-157). Springer.

[38]. Ghadge, A., Kidd, E., Bhattacharjee, A., & Tiwari, M. K. (2019). Sustainable procurement performance of large enterprises across supply chain tiers and geographic regions. *International journal of production research*, *57*(3), 764-778.

[39]. Ghazali, E. M., Nguyen, B., Mutum, D. S., & Yap, S.-F. (2019). Pro-environmental behaviours and Value-Belief-Norm theory: Assessing unobserved heterogeneity of two ethnic groups. *Sustainability*, *11*(12), 3237.

[40]. Ghazi, A., & Lotfi, F. H. (2019). Assessment and budget allocation of Iranian natural gas distribution company-A CSW DEA based model. *Socio-economic planning sciences*, *66*, 112-118.

[41]. Ghisetti, C. (2017). Demand-pull and environmental innovations: Estimating the effects of innovative public procurement. *Technological forecasting and social change*, *125*, 178-187.

[42]. Golas, S. B., Shibahara, T., Agboola, S., Otaki, H., Sato, J., Nakae, T., Hisamitsu, T., Kojima, G., Felsted, J., & Kakarmath, S. (2018). A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. *BMC medical informatics and decision making*, *18*(1), 44.

[43]. Górecki, J., & Díaz-Madroñero, M. (2020). Who risks and wins?—Simulated cost variance in sustainable construction projects. *Sustainability*, *12*(8), 3370.

[44]. Hamari, J., & Keronen, L. (2017). Why do people buy virtual goods: A meta-analysis. *Computers in Human Behavior*, *71*, 59-69.

[45]. Han, Y., Chong, W. K., & Li, D. (2020). A systematic literature review of the capabilities and performance metrics of supply chain resilience. *International journal of production research*, *58*(15), 4541-4566.

[46]. Haseeb, A., Xia, E., Saud, S., Ahmad, A., & Khurshid, H. (2019). Does information and communication technologies improve environmental quality in the era of globalization? An empirical analysis. *Environmental Science and Pollution Research*, *26*(9), 8594-8608.

[47]. Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., & Vazire, S. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181-188.

[48]. Hu, M. (2019). Does zero energy building cost more?–An empirical comparison of the construction costs for zero energy education building in United States. *Sustainable cities and society*, *45*, 324-334.

[49]. Inderfurth, K., Kelle, P., & Kleber, R. (2018). Inventory control in dual sourcing commodity procurement with price correlation. *Central European Journal of Operations Research*, *26*(1), 93-119.

[50]. Ivanov, D., Dolgui, A., Sokolov, B., & Ivanova, M. (2017). Literature review on disruption recovery in the supply chain. *International journal of production research*, *55*(20), 6158-6174.

[51]. Jensen, J. D., Christensen, T., Denver, S., Ditlevsen, K., Lassen, J., & Teuber, R. (2019). Heterogeneity in consumers' perceptions and demand for local (organic) food products. *Food quality and preference*, *73*, 255-265.

[52]. Jin, H., Shen, L., & Wang, Z. (2018). Mapping the influence of project management on project cost. *KSCE Journal of Civil Engineering*, *22*(9), 3183-3195.

[53]. Ju, H., & Xu, S. (2017). Research status of earned value management. Proceedings of the Fourth International Forum on Decision Sciences,

[54]. Kanakaris, N., Karacapilidis, N., Kournetas, G., & Lazanas, A. (2019). Combining machine learning and operations research methods to advance the project management practice. International Conference on Operations Research and Enterprise Systems,

[55]. Kavuma, A., Ock, J., & Jang, H. (2019). Factors influencing time and cost overruns on freeform construction projects. *KSCE Journal of Civil Engineering*, *23*(4), 1442-1450.

[56]. Khan, H., Khan, I., & Binh, T. T. (2020). The heterogeneity of renewable energy consumption, carbon emission and financial development in the globe: A panel quantile regression approach. *Energy Reports*, *6*, 859-867.

[57]. Khan, I., Hou, F., Zakari, A., & Tawiah, V. K. (2021). The dynamic links among energy transitions, energy consumption, and sustainable economic growth: A novel framework for IEA countries. *Energy*, *222*, 119935.

[58]. Khodeir, L. M., & El Ghandour, A. (2019). Examining the role of value management in controlling cost overrun [application on residential construction projects in Egypt]. *Ain Shams Engineering Journal*, *10*(3), 471-479.

[59]. Khoshbakht, M., Gou, Z., & Dupre, K. (2017). Cost-benefit prediction of green buildings: SWOT analysis of research methods and recent applications. *Procedia Engineering*, *180*, 167-178.

[60]. Kim, B.-C., & Kwak, Y. H. (2018). Improving the accuracy and operational predictability of project cost forecasts: an adaptive combination approach. *Production Planning & Control*, *29*(9), 743-760.

[61]. Koch, C. (2019). Machine learning for engineering processes. International Conference on Business Information Systems,

[62]. Kuo, R.-J., Nugroho, Y., & Zulvia, F. E. (2019). Application of particle swarm optimization algorithm for adjusting project contingencies and response strategies under budgetary constraints. *Computers & Industrial Engineering*, *135*, 254-264.

[63]. Lake, D. R. (2019). The Shrinking American Military. In *The Pursuit of Technological Superiority and the Shrinking American Military* (pp. 17-62). Springer.
[64]. Lane, S. R., & Pritzker, S. (2017). Budgeting and Allocating Resources. *Political Social Work: Using Power to Create Social Change*, 369-397.
[65]. Laurent, C. S., & Cowlagi, R. V. (2021a). Coupled sensor configuration and path-planning in unknown static environments. 2021 American Control Conference (ACC),
[66]. Laurent, C. S., & Cowlagi, R. V. (2021b). Depth-first coupled sensor configuration and path-planning in unknown static environments. 2021 European Control Conference (ECC),
[67]. Laureti, T., & Benedetti, I. (2018). Exploring pro-environmental food purchasing behaviour: An empirical analysis of Italian consumers. *Journal of cleaner production*, *172*, 3367-3378.
[68]. Lăzăroiu, G., Ionescu, L., Uță, C., Hurloiu, I., Andronie, M., & Dijmărescu, I. (2020). Environmentally responsible behavior and sustainability policy adoption in green public procurement. *Sustainability*, *12*(5), 2110.
[69]. Le, T.-A., Huynh, Q.-T., Nguyen, T.-H., Nguyen, N.-H., & Cao, P.-N. (2020). A method for project completion cost predicting using lstm in earned value management technique. 2020 4th International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom),
[70]. Li, R., Chen, Y., Ritchie, M. D., & Moore, J. H. (2020). Electronic health records and polygenic risk scores for predicting disease risk. *Nature Reviews Genetics*, *21*(8), 493-502.
[71]. Liggett, W., Hunter, H., & Jones, M. (2019). So you passed an earned value management government validation-now what? 2019 IEEE Aerospace Conference,
[72]. Liu, J., Capurro, D., Nguyen, A., & Verspoor, K. (2021). Early prediction of diagnostic-related groups and estimation of hospital cost by processing clinical notes. *NPJ digital medicine*, *4*(1), 103.
[73]. López Pascual, J., Meléndez Rodríguez, J. C., & Cruz Rambaud, S. (2021). The Enhanced-Earned Value Management (E-EVM) model: A proposal for the aerospace industry. *Symmetry*, *13*(2), 232.
[74]. Love, P. E., & Ahiaga-Dagbui, D. D. (2018). Debunking fake news in a post-truth era: The plausible untruths of cost underestimation in transport infrastructure projects. *Transportation research part A: policy and practice*, *113*, 357-368.
[75]. Love, P. E., Ika, L. A., Matthews, J., & Fang, W. (2021). Risk and uncertainty in the cost contingency of transport projects: Accommodating bias or heuristics, or both? *IEEE transactions on engineering management*, *71*, 205-219.
[76]. Mansura Akter, E., & Sumaiya, M. (2022). Antimicrobial Resistance Profiles of Acinetobacter Baumannii and Vibrio Cholerae in Hospital Wastewater and Adjacent Household Water Sources. *American Journal of Data Science and Analytics*, *3*(09), 01-55. https://doi.org/10.63125/gbervt14
[77]. Martens, A., & Vanhoucke, M. (2017). A buffer control method for top-down project control. *European Journal of Operational Research*, *262*(1), 274-286.
[78]. Massey, M., O'Cass, A., & Otahal, P. (2018). A meta-analytic study of the factors driving the purchase of organic food. *Appetite*, *125*, 418-427.
[79]. Md Imran, H., & Jannatul, F. (2022). Molecular and Precision Rehabilitation Molecular Mechanisms Linking Mitochondrial Dysfunction, Inflammation, and Muscle Degeneration with Functional Decline and Rehabilitation Response in Neuromuscular Disease. *American Journal of Interdisciplinary Studies*, *12*(08), 01-44. https://doi.org/10.63125/0ym5bh79
[80]. Md Maksudul, H., & Md Tazul Islam, C. (2022). Machine Learning-Based Cash-Flow Forecasting for Improving Financial Planning in U.S. Small and Mid-Sized Businesses. *International Journal of Scientific Interdisciplinary Research*, *12*(12), 01–38. https://doi.org/10.63125/w3pgjx97
[81]. Md Shahid, N. (2023). Digital Twin-Based Network Reliability Modeling for LAN And WAN Infrastructure Resilience. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 577–629. https://doi.org/10.63125/ahjvxj14
[82]. Md Tazul Islam, C., & Ahmed, H. C. H. R. (2023). Predictive Operations Analytics for Resource Allocation and Service-Capacity Planning in Data-Driven Organizations. *Review of Applied Science and Technology*, *13*(02), 01–41. https://doi.org/10.63125/z3sm3089
[83]. Md. Morshedul, I., & Aminul Islam, M. (2022). A Quantitative Evaluation of Explainable Ensemble Learning for Probabilistic Risk Estimation and Decision Reliability in Enterprise Analytics. *American Journal of Data Science and Analytics*, *3*(02), 01-38. https://doi.org/10.63125/k22shk74
[84]. Milat, M., Knezić, S., & Sedlar, J. (2021). Resilient scheduling as a response to uncertainty in construction projects. *Applied Sciences*, *11*(14), 6493.
[85]. Miller, G. J., Robbins, D., & Keum, J. (2017). Conventional Budgeting with Targets, Incentives, and Performance. In *Government Budgeting and Financial Management in Practice* (pp. 123-148). Routledge.
[86]. Mills, M., & Kanavos, P. (2020). Do pharmaceutical budgets deliver financial sustainability in healthcare? Evidence from Europe. *Health Policy*, *124*(3), 239-251.
[87]. Mir, M., Kabir, H. D., Nasirzadeh, F., & Khosravi, A. (2021). Neural network-based interval forecasting of construction material prices. *Journal of Building Engineering*, *39*, 102288.
[88]. Mohammed, K. N., & Chambrelin, K. S. (2020). An analytical approach in usage of agile methodologies in construction industries–A case study. *Materials Today: Proceedings*, *33*, 475-479.
[89]. Moons, K., Waeyenbergh, G., & Pintelon, L. (2019). Measuring the logistics performance of internal hospital supply chains–a literature study. *Omega*, *82*, 205-217.
[90]. Morgan, D. F., Robinson, K. S., Strachota, D., & Hough, J. A. (2017). Performance Budgeting. In *Budgeting for Local Governments and Communities* (pp. 345-376). Routledge.
[91]. Mühlpfordt, T., Roald, L., Hagenmeyer, V., Faulwasser, T., & Misra, S. (2019). Chance-constrained AC optimal power flow: A polynomial chaos approach. *IEEE Transactions on Power Systems*, *34*(6), 4806-4816.

[92]. Nisa, C. F., Bélanger, J. J., Schumpe, B. M., & Faller, D. G. (2019). Meta-analysis of randomised controlled trials testing behavioural interventions to promote household action on climate change. *Nature communications*, *10*(1), 4545.
[93]. Nuttavuthisit, K., & Thøgersen, J. (2017). The importance of consumer trust for the emergence of a market for green products: The case of organic food. *Journal of business ethics*, *140*(2), 323-337.
[94]. Osojnik, A., Panov, P., & Džeroski, S. (2018). Tree-based methods for online multi-target regression. *Journal of Intelligent Information Systems*, *50*(2), 315-339.
[95]. Ou-Yang, C., & Chen, W.-L. (2017). Applying a risk assessment approach for cost analysis and decision-making: a case study for a basic design engineering project. *Journal of the Chinese Institute of Engineers*, *40*(5), 378-390.
[96]. Ouyang, X., Li, Q., & Du, K. (2020). How does environmental regulation promote technological innovations in the industrial sector? Evidence from Chinese provincial panel data. *Energy policy*, *139*, 111310.
[97]. Pal, R., Wang, P., & Liang, X. (2017). The critical factors in managing relationships in international engineering, procurement, and construction (IEPC) projects of Chinese organizations. *International journal of project management*, *35*(7), 1225-1237.
[98]. Pan, Y., Boutselis, G. I., & Theodorou, E. A. (2018). Efficient reinforcement learning via probabilistic trajectory optimization. *IEEE transactions on neural networks and learning systems*, *29*(11), 5459-5474.
[99]. Pan, Y., Saigol, K., & Theodorou, E. A. (2017). Belief space stochastic control under unknown dynamics. 2017 American Control Conference (ACC),
[100]. Paul, A., Kamwa, I., & Joos, G. (2018). Centralized dynamic state estimation using a federation of extended Kalman filters with intermittent PMU data from generator terminals. *IEEE Transactions on Power Systems*, *33*(6), 6109-6119.
[101]. Queiroz, M. M., Fosso Wamba, S., De Bourmont, M., & Telles, R. (2021). Blockchain adoption in operations and supply chain management: empirical evidence from an emerging economy. *International journal of production research*, *59*(20), 6087-6103.
[102]. Rajeev, A., Pati, R. K., Padhi, S. S., & Govindan, K. (2017). Evolution of sustainability in supply chain management: A literature review. *Journal of cleaner production*, *162*, 299-314.
[103]. Ranta, V., Aarikka-Stenroos, L., & Mäkinen, S. J. (2018). Creating value in the circular economy: A structured multiple-case analysis of business models. *Journal of cleaner production*, *201*, 988-1000.
[104]. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, *4*(1), 86.
[105]. Rendall, R., Pereira, A. C., & Reis, M. S. (2017). Advanced predictive methods for wine age prediction: Part I–A comparison study of single-block regression approaches based on variable selection, penalized regression, latent variables and tree-based ensemble methods. *Talanta*, *171*, 341-350.
[106]. Ruparathna, R., Hewage, K., & Sadiq, R. (2017). Economic evaluation of building energy retrofits: A fuzzy based approach. *Energy and Buildings*, *139*, 395-406.
[107]. Sackey, S., Lee, D.-E., & Kim, B.-S. (2020). Duration estimate at completion: Improving earned value management forecasting accuracy. *KSCE Journal of Civil Engineering*, *24*(3), 693-702.
[108]. Sadia, A., & Md. Mosheur, R. (2023). Consumer Trust, Perceived Authenticity, and Personalized Digital Marketing: A Quantitative Decision-Science Study. *International Journal of Scientific Interdisciplinary Research*, *4*(2), 252–295. https://doi.org/10.63125/cajkqt60
[109]. Sánchez-Flores, R. B., Cruz-Sotelo, S. E., Ojeda-Benitez, S., & Ramírez-Barreto, M. E. (2020). Sustainable supply chain management—A literature review on emerging economies. *Sustainability*, *12*(17), 6972.
[110]. Sassani, A., Smadi, O., & Hawkins, N. (2021). Developing pavement marking management systems: A theoretical model framework based on the experiences of the US transportation agencies. *Infrastructures*, *6*(2), 18.
[111]. Saura, J. R., Palos-Sánchez, P., & Cerdá Suárez, L. M. (2017). Understanding the digital marketing environment with KPIs and web analytics. *Future internet*, *9*(4), 76.
[112]. Schuster, P., Heinemann, M., & Cleary, P. (2021). Variance Analysis and Control. In *Management Accounting* (pp. 173-214). Springer.
[113]. Scott Morton, F. M., Stern, A. D., & Stern, S. (2018). The impact of the entry of biosimilars: evidence from Europe. *Review of Industrial Organization*, *53*(1), 173-210.
[114]. Shah, M. N., Dixit, S., Kumar, R., Jain, R., & Anand, K. (2021). Causes of delays in slum reconstruction projects in India. *International Journal of Construction Management*, *21*(5), 452-467.
[115]. Sharif, A., Raza, S. A., Ozturk, I., & Afshan, S. (2019). The dynamic relationship of renewable and nonrenewable energy consumption with carbon emission: a global study with the application of heterogeneous panel estimations. *Renewable energy*, *133*, 685-691.
[116]. Shehadeh, A., Alshboul, O., Al Mamlook, R. E., & Hamedat, O. (2021). Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. *Automation in Construction*, *129*, 103827.
[117]. Sjödin, D., Kamalaldin, A., Parida, V., & Islam, N. (2021). Procurement 4.0: How industrial customers transform procurement processes to capitalize on digital servitization. *IEEE transactions on engineering management*, *70*(12), 4175-4190.
[118]. Small, E. P., Bakry, I., & Ayyash, L. a. (2021). Evaluating the effect of TQM on MEP construction productivity and project delivery in Dubai. *International Journal of Construction Management*, *21*(10), 1061-1075.
[119]. Song, C., Pons, A., & Yen, K. (2021). AG-SGD: angle-based stochastic gradient descent. *IEEE Access*, *9*, 23007-23024.
[120]. Stabak, D., & Md Shahid, N. (2022). Empirical Assessment of Uncertainty-Aware Real-Time Decision-Making for Reliable and Resilient Cyber-Physical Systems. *International Journal of Scientific Interdisciplinary Research*, *1*(01), 432–477. https://doi.org/10.63125/q3r4kk16

[121]. Steward, M. D., Narus, J. A., Roehm, M. L., & Ritz, W. (2019). From transactions to journeys and beyond: The evolution of B2B buying process modeling. *Industrial Marketing Management*, *83*, 288-300.
[122]. Syed Feroz, A., & Md Mahamud Pervaz, P. (2022). Root-Cause Analysis and FMEA-Driven Reliability Improvement in Gas Production Facilities: A Multi-Year Empirical Study. *International Journal of Scientific Interdisciplinary Research*, *12*(10), 01–53. https://doi.org/10.63125/70msp878
[123]. Syed Feroz, A., & Shabbir, A. (2022). Finite Element-Based Thermomechanical Optimization of Pressure Components for Extending Industrial Boiler Service Life. *Journal of Sustainable Development and Policy*, *12*(12), 01-51. https://doi.org/10.63125/x24pht28
[124]. Taherdoost, H., & Brard, A. (2019). Analyzing the process of supplier selection criteria and methods. *Procedia manufacturing*, *32*, 1024-1034.
[125]. Tang, B.-J., Wang, X.-Y., & Wei, Y.-M. (2019). Quantities versus prices for best social welfare in carbon reduction: A literature review. *Applied energy*, *233*, 554-564.
[126]. Täuscher, K., & Laudien, S. M. (2018). Understanding platform business models: A mixed methods study of marketplaces. *European management journal*, *36*(3), 319-329.
[127]. Topcu, M., & Tugcu, C. T. (2020). The impact of renewable energy consumption on income inequality: Evidence from developed countries. *Renewable energy*, *151*, 1134-1140.
[128]. Tyagi, M. K., Sikandar, A., Tyagi, D. K., Kumar, D., Singh, P., Munisamy, S., & Reddy, L. (2020). State space modeling of Earned Value Method for iterative enhancement based traditional software projects tracking. International Advanced Computing Conference,
[129]. Ulusoy, G., & Hazır, Ö. (2021). Progress and Cost Control. In *An Introduction to Project Modeling and Planning* (pp. 311-334). Springer.
[130]. Vanaei, H., Eslami, A., & Egbewande, A. (2017). A review on pipeline corrosion, in-line inspection (ILI), and corrosion growth rate models. *International Journal of Pressure Vessels and Piping*, *149*, 43-54.
[131]. Vrablecová, P., Ezzeddine, A. B., Rozinajová, V., Šárik, S., & Sangaiah, A. K. (2018). Smart grid load forecasting using online support vector regression. *Computers & Electrical Engineering*, *65*, 102-117.
[132]. Vrontis, D., Thrassou, A., Santoro, G., & Papa, A. (2017). Ambidexterity, external knowledge and performance in knowledge-intensive firms. *The Journal of Technology Transfer*, *42*(2), 374-388.
[133]. Walter, A.-T. (2021). Organizational agility: ill-defined and somewhat confusing? A systematic literature review and conceptualization. *Management Review Quarterly*, *71*(2), 343-391.
[134]. Wang, H., Xiong, W., Wu, G., & Zhu, D. (2018). Public–private partnership in Public Administration discipline: a literature review. *Public management review*, *20*(2), 293-316.
[135]. Wang, X.-W., Cao, Y.-M., & Zhang, N. (2021). The influences of incentive policy perceptions and consumer social attributes on battery electric vehicle purchase intentions. *Energy policy*, *151*, 112163.