Quanta Research
https://journals.tultech.eu/index.php/qr
ISSN: 2806-3279
Volume 4, Issue 1
DOI: https://doi.org/10.15157/qr.2026.4.1.181-197
Received: 12.07.2026; Revised: 10.09.2029; Accepted: 01.10.2026

Research Article

Machine Learning-Driven Cost Estimation and Resource Allocation Optimization in Industrial Engineering Projects: Evidence from Albania

Fiona Shabani[1*], Ermal Osmani[1], Amr Badr [2]

[1]Department of Mechanics, Polytechnic University of Tirana, Tirana, Albania
[2]School of Science and Technology, University of New England, Armidale, Australia
**fiona.shabani@fim.edu.al*

Abstract

Cost overruns and inefficient resource allocation remain persistent problems in industrial engineering projects, particularly in transition economies where historical data are scarce and price volatility is high. This study develops an integrated predict-then-optimize framework that couples machine learning (ML) cost estimation with risk-aware resource allocation, and demonstrates it on a calibrated case study of 420 industrial projects (manufacturing plants, logistics facilities, agro-food processing, energy, water treatment and mineral processing) in eight Albanian regions over 2015–2025. Five learners' multiple linear regression, support vector regression, random forest, artificial neural network and gradient boosting were trained in a direct formulation and in a hybrid formulation that learns the deviation from a conventional unit-rate estimate indexed to the national construction cost index. Under a strict temporal hold-out (training 2015–2022, testing 2023–2025), the hybrid gradient boosting model reduced the mean absolute percentage error from 19.1% for the conventional estimate to 8.2% (R^2 = 0.974), whereas direct tree-based models degraded under post-2021 price escalation. Design completeness at tender, site difficulty and region were the dominant cost-risk drivers. Conformalised quantile regression provided P10–P90 cost intervals that were fed into a mixed-integer linear programme selecting execution modes and start times for a 14-project portfolio under three resource capacities. In Monte-Carlo evaluation the risk-aware schedule lowered expected portfolio cost by 3.1% (≈ EUR 2.4 million), reduced the expected number of late projects by 26% and cut expected delay-months by 67% relative to earliest-due-date practice, with the largest gains when resources were scarce. The framework offers project-based firms and public clients in Albania a transparent, reproducible decision tool.

Keywords: Cost estimation; Machine learning; Resource allocation; Mixed-integer programming; Conformal prediction; Albania.

INTRODUCTION

Industrial engineering projects of the construction, expansion or retrofit for manufacturing plants, warehouses, processing facilities, energy installations and treatment plants are the physical channel through which investment becomes productive capacity. Their economic performance depends on two decisions taken early and under considerable

© 2026 Authors. This is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International License CC BY 4.0 (http://creativecommons.org/licenses/by/4.0).

uncertainty: how much the project will cost, and how scarce engineering staff, skilled crews and equipment should be allocated across concurrent projects. International evidence shows that both decisions are frequently made poorly. Large samples of infrastructure and industrial projects reveal systematic cost underestimation, with average overruns of 20–45% depending on project type [1,2], and the debate on whether overruns stem from technical error, optimism bias or strategic misrepresentation remains open [3,4].

These problems are amplified in Albania. The economy has experienced sustained investment in industrial parks, agro-processing, renewable energy and water infrastructure, yet contractors and public clients operate with fragmented cost records, short time series and pronounced price shocks. The construction cost index for dwellings published by the Institute of Statistics (INSTAT) rose to 113.5 in the second quarter of 2025 relative to its fourth-quarter 2021 base [5], and wage growth in construction has been persistent although construction wages remain below the national average [6]. Conventional estimating practice — bills of quantities priced with published unit rates and escalated with a price index — cannot capture how design maturity, procurement route, competition, site conditions and contractor capability modify the final cost. At the same time, regional engineering-procurement-construction (EPC) firms typically schedule portfolios with simple priority rules, so that resource conflicts are resolved reactively.

Machine learning (ML) offers a way to learn these non-linear effects from historical projects, and a growing literature reports that ML models outperform parametric regression in construction cost prediction [7,8]. Separately, operations research provides mature models for resource allocation and resource-constrained project scheduling [9,10]. The two streams, however, are rarely connected: predictions are evaluated by accuracy metrics alone, while scheduling models assume costs are known constants. Recent work on prescriptive analytics argues that predictions should be judged by the quality of the decisions they enable [11,12]. This study addresses that gap for industrial engineering projects in a data-scarce transition economy.

The contributions are fourfold. First, a hybrid ML formulation is proposed in which learners model the deviation of actual cost from an index-adjusted unit-rate estimate, and it is shown to be markedly more robust to price-level drift than direct ML estimation. Second, conformalised quantile regression is used to attach calibrated P10–P90 intervals to each estimate. Third, the predicted cost distributions feed a mixed-integer linear programme (MILP) that jointly selects execution modes and start times for a project portfolio under multiple resource capacities, delay penalties and escalation. Fourth, the full pipeline is demonstrated and evaluated on a calibrated Albanian case study with Monte-Carlo validation and sensitivity analysis, and all code is made available for replication.

The remainder of the paper is organised as follows. The next section reviews related work. The methodology section describes the case-study data, the estimation models, the uncertainty quantification and the optimization model. Results are then presented and discussed, followed by conclusions, limitations and directions for further research.

LITERATURE REVIEW

Machine Learning for Cost Estimation

Early applications of artificial neural networks (ANN) to parametric cost estimation of highway projects [13] and building projects [14] demonstrated that non-linear models could reduce estimating error relative to linear regression. Kim et al. [15] compared regression, ANN and case-based reasoning on 530 residential projects and found the ANN most accurate, while Cheng et al. [16] combined fuzzy logic, neural networks and evolutionary optimization for conceptual estimates. More recently, Matel et al. [17] estimated engineering-service costs with ANN, and systematic reviews [7,8] identify support vector regression (SVR) [18], random forests [19] and gradient boosting machines [20,21] as the most frequently successful learners. Benchmarks on tabular data generally confirm that tree ensembles outperform deep networks on small and medium data sets [22,23]. Two weaknesses recur in this literature: random cross-validation, which overstates accuracy when projects from the same period appear in training and test sets, and the absence of explicit treatment of price-level change, which tree-based models cannot extrapolate.

Uncertainty Quantification

A single-point estimate is of limited value for budgeting, contingency setting or risk-aware scheduling. Quantile regression [24] and quantile regression forests [25] estimate conditional quantiles directly, but their empirical coverage is not guaranteed in finite samples. Conformal prediction [26] provides distribution-free coverage guarantees under exchangeability, and conformalised quantile regression (CQR) [27] combines the adaptivity of quantile models with conformal calibration. Despite its attractiveness for cost contingency estimation, CQR has seen little use in engineering project management.

Resource Allocation and Project Scheduling

Multi-project scheduling with limited resources was formulated as a 0–1 programme by Pritsker et al. [28]; the resource-constrained project scheduling problem (RCPSP) is strongly NP-hard [29] and has inspired a large body of exact and heuristic methods [10]. In construction, genetic algorithms have been applied to resource levelling [9] and to the time–cost trade-off problem [30]. Herroelen and Leus [31] review proactive and reactive scheduling under uncertainty, and risk measures such as conditional value-at-risk (CVaR) [32] allow decision-makers to penalise unfavourable outcomes. Few studies, however, let ML-predicted cost distributions drive the cost coefficients of the allocation model.

Research Gap

The review identifies three gaps: (i) ML cost models are seldom tested under temporal hold-out conditions that mimic real use during price shocks; (ii) predictive uncertainty is rarely calibrated and almost never propagated into resource decisions; and (iii) evidence from Western Balkan economies, including Albania, is scarce. The present study responds to all three.

MATERIALS AND METHODS

The research follows the predict-then-optimize framework summarised in Figure 1. A project database is assembled and pre-processed; cost models are trained and validated; calibrated prediction intervals are produced; and the resulting cost distributions parameterise a resource allocation MILP whose decisions are evaluated by Monte-Carlo simulation.

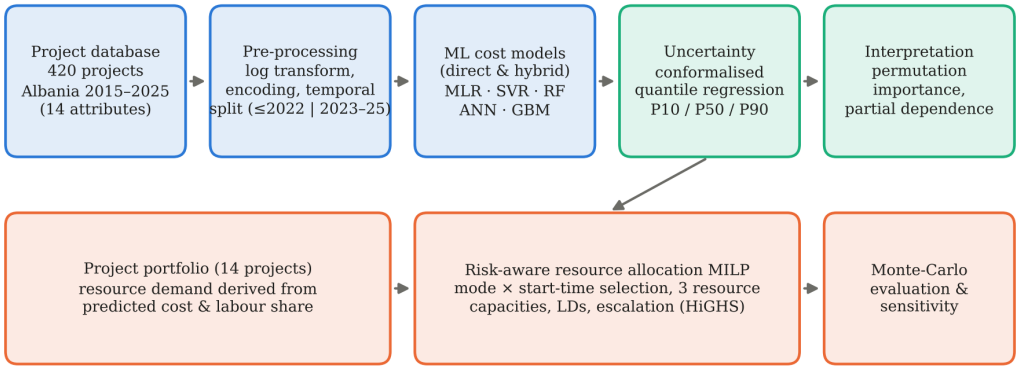


Figure 1. Integrated predict-then-optimize framework for cost estimation and resource allocation

Case Study Context and Data

Project-level cost records of Albanian industrial projects are not publicly available in a consolidated form: contract notices published through the Public Procurement Agency [33] report contract values but not final costs, and private projects are not reported at all. To provide a transparent and fully reproducible test bed, the study therefore uses a calibrated simulation case study — a structured 'digital twin' of the project portfolio of a regional EPC market. The database contains $n = 420$ industrial projects started between 2015 and 2025 in eight counties (qarqe): Tiranë, Durrës, Elbasan, Fier, Vlorë, Shkodër, Korçë and Dibër. Six project types are represented: manufacturing plants (94), warehouse and logistics facilities (90), agro-food processing plants (78), energy projects such as photovoltaic parks and small hydropower plants (64), mining and mineral processing facilities (53), and water and wastewater treatment plants (41).

Each project is described by quantity drivers (gross floor area, structural steel, concrete volume and installed mechanical-electrical-plumbing (MEP) power), schedule and context variables (planned duration, design completeness at tender, contractor experience, number of bidders, site difficulty on a five-point scale), categorical attributes (project type, region, procurement method and funding source) and the construction cost index (CCI) at start. The CCI profile is anchored to INSTAT's base of Q4-2021 = 100 and its Q2-2025 value of 113.5 [5], with the strong 2022 escalation reflected in the series. Regional price factors, labour shares and resource intensities were set to plausible values for the Albanian market, and the actual cost of each project was generated by a documented multiplicative process in which design immaturity, weak competition, negotiated procurement, public funding,

difficult sites and schedule compression increase cost relative to the conventional estimate, with lognormal residual noise ($\sigma = 0.085$). Because the generating process is known, the realised cost of any alternative decision can be simulated, which is exactly what is required to evaluate the optimization stage without hindsight bias. Descriptive statistics are given in Table 1.

Table 1. Descriptive statistics of the numerical variables ($n = 420$)

Variable	Unit	Mean	SD	Min	Median	Max
Gross floor area A	m²	5,907	4,167	640	4,830	30,660
Structural steel	t	300	238	32	221	1,722
Concrete volume	m³	3,033	2,099	314	2,480	12,538
Installed MEP power	kW	1,625	2,040	91	960	16,134
Planned duration	months	8.9	3.1	4.0	8.0	23.0
Design completeness	%	68.3	15.7	35.0	70.5	98.0
Contractor experience	years	11.8	7.1	1.0	11.0	35.0
Number of bidders	–	4.3	2.4	1.0	4.0	12.0
Site difficulty	1–5	2.81	1.27	1.00	3.00	5.00
Construction cost index	Q4-2021=100	101.7	8.2	90.1	98.0	115.0
Conventional estimate	M EUR	3.58	2.26	0.52	2.96	15.39
Actual cost	M EUR	4.30	2.79	0.43	3.55	18.54
Overrun vs. estimate	%	21.4	25.6	-21.5	16.7	132.2

Measured against the conventional estimate, the average overrun is 21.4% (median 16.7%); 80% of projects exceed the estimate and 62% exceed it by more than 10%. These magnitudes are consistent with international evidence [1,2]. Figure 2(a) shows that the dispersion of overruns is wide in every project type, while Figure 2(b) confirms the expected scale relationship between floor area and cost, with considerable scatter attributable to non-quantity drivers.

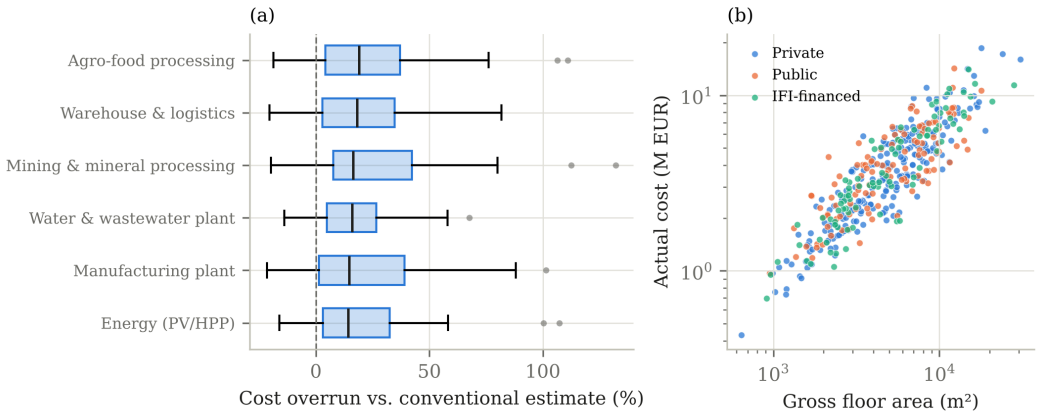


Figure 2. (a) Distribution of cost overrun relative to the conventional estimate by project type; (b) actual cost versus gross floor area by funding source (log scales)

Cost Estimation Models

The benchmark is the conventional deterministic estimate used in practice, in which quantities are priced at published unit rates and escalated by the CCI:

$$\hat{C}_i^{\text{conv}} = \frac{I_{t_i}}{100}\left(A_i u_{k(i)} + \sum_{j\in\{s,c,e\}} q_{ij}\, u_j \right) \tag{1}$$

where A_i is the floor area priced at the type-specific shell rate $u_{k(i)}$, q_{ij} are steel, concrete and MEP quantities with unit rates u_j, and I_t is the cost index at start. Two ML formulations are compared. The *direct* formulation learns the log-cost as a function of the feature vector $\mathbf{x}_i$:

$$\ln C_i = f(\mathbf{x}_i) + \varepsilon_i \tag{2}$$

whereas the *hybrid* formulation uses Eq. (1) as an offset and learns only the log-deviation $g(\mathbf{x})$, so that price-level and quantity effects are carried by engineering knowledge and the learner concentrates on risk factors:

$$\ln C_i = \ln \hat{C}_i^{\text{conv}} + g(\mathbf{x}_i) + \varepsilon_i, \qquad \hat{C}_i = \hat{C}_i^{\text{conv}} \exp\big(\hat{g}(\mathbf{x}_i)\big) \tag{3}$$

Five learners are fitted in both formulations: multiple linear regression on log-transformed quantities (MLR), SVR with a radial basis kernel [18], random forest (RF) [19], a multilayer perceptron (ANN) and gradient boosting (GBM) [20]. Gradient boosting builds an additive ensemble of shallow regression trees by stagewise fitting of the negative gradient of the loss [20]:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \sum_{l=1}^{L_m} \gamma_{lm}\, \mathbb{1}(\mathbf{x} \in R_{lm}) \tag{4}$$

where ν is the learning rate and γ_{lm} are the leaf values of the m-th tree with regions R_{lm}. Categorical variables are one-hot encoded and numerical variables standardised, with logarithms of the quantity drivers added as features [34]. Hyper-parameters are selected by grid search with 10-fold cross-validation on the training set. To reflect real use — estimating new projects with models trained on the past — the evaluation uses a strict temporal split [35]: projects started in 2015–2022 form the training set (282 projects) and those started in 2023–2025 form the test set (138 projects). Accuracy is reported as mean and median absolute percentage error (MAPE, MdAPE), root mean squared error (RMSE), the coefficient of determination and the share of projects estimated within ±10%:

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n} \frac{|\hat{C}_i - C_i|}{C_i}, \qquad R^2 = 1 - \frac{\sum_i (C_i - \hat{C}_i)^2}{\sum_i (C_i - \bar{C})^2} \tag{5}$$

Interpretation uses permutation importance [36], measuring the increase in test error when a feature is randomly permuted, and partial dependence functions [20] showing the average marginal effect of a feature; both are model-agnostic and complement Shapley-value approaches [37].

Uncertainty Quantification

Quantile gradient boosting models for $\tau = 0.1$ and $\tau = 0.9$ are trained by minimising the pinball loss [24]:

$$\rho_\tau(u) = u(\tau - \mathbb{1}\{u < 0\}), \qquad \hat{q}_\tau = \operatorname{argmin}_q \sum_i \rho_\tau\left(y_i - q(\mathbf{x}_i)\right) \tag{6}$$

Following the CQR procedure [27], the quantile models are first fitted on projects started up to 2020, and conformity scores are computed on the 2021–2022 calibration projects:

$$E_i = \max\{\hat{q}_{0.1}(\mathbf{x}_i) - y_i,\ y_i - \hat{q}_{0.9}(\mathbf{x}_i)\}, \qquad \hat{I}(\mathbf{x}) = [\hat{q}_{0.1}(\mathbf{x}) - Q_{0.8}(E),\ \hat{q}_{0.9}(\mathbf{x}) + Q_{0.8}(E)] \tag{7}$$

where $Q_{0.8}(E)$ is the empirical $(1-\alpha)(1+1/n_{\text{cal}})$ quantile of the scores. The models are then refitted on all training data and the calibration offset is retained. To guarantee coherent distributions, the quantiles are rearranged so that P10 ≤ P50 ≤ P90 [38], where P50 is the hybrid GBM point estimate.

Risk-Aware Resource Allocation Model

The optimization stage considers a regional EPC contractor launching a portfolio P of 14 projects drawn from the 2023–2025 test set, over a planning horizon of H = 34 months. Each project can be executed in one of three modes $m \in M$: accelerated (80% of the normal duration, with 10% more total effort), normal, or extended (125% of the normal duration, with 3% less effort). For every project and mode, the ML models supply the predicted cost distribution $\{\hat{C}^{10}_{pm}, \hat{C}^{50}_{pm}, \hat{C}^{90}_{pm}\}$ by changing the planned duration feature, so that the time–cost trade-off is learned from data rather than assumed. Three renewable resources $k \in K$ are considered — engineers, skilled crews of six workers, and heavy equipment units — with monthly demand

$$r_{pmk} = \frac{\eta_m\, s_{pk}\, \hat{C}^{50}_{pm}}{w_k\, d_{pm}} \tag{8}$$

where η_m is the mode effort multiplier, s_{pk} the share of cost absorbed by resource k in the project's type, w_k the monthly unit cost of the resource (calibrated to Albanian wage levels [6]) and d_{pm} the duration. The binary variable x_{pmt} equals 1 if project p starts in month t in mode m. The model minimises total risk-adjusted cost:

$$\min_{\mathbf{x}} Z = \sum_{p\in P}\sum_{m\in M}\sum_{t\in T_{pm}} c_{pmt}\, x_{pmt} \tag{9}$$

$$c_{pmt} = \hat{C}^{50}_{pm}(1+e)^t + \lambda\left(\hat{C}^{90}_{pm} - \hat{C}^{50}_{pm}\right) + \min\{\kappa,\ \delta\,\Theta_{pmt}(\alpha)\}\hat{C}^{50}_{pm} \tag{10}$$

$$\Theta_{pmt}(\alpha) = \max\{0,\ t + d_{pm}\, e^{\mu + z_\alpha \sigma_{pm}} - D_p\} \tag{11}$$

subject to

$$\sum_{m\in M}\sum_{t\in T_{pm}} x_{pmt} = 1 \qquad \forall p \in P \tag{12}$$

$$\sum_{p\in P}\sum_{m\in M}\sum_{t=\tau-d_{pm}+1}^{\tau} r_{pmk}\, x_{pmt} \le R_k \qquad \forall k \in K,\ \tau = 0,\dots,H-1 \tag{13}$$

$$x_{pmt} \in \{0,1\}, \qquad T_{pm} = \left\{t : \rho_p \le t,\ t + d_{pm} \le \min(H,\ D_p + \Delta)\right\} \tag{14}$$

The cost coefficient (10) combines the median cost escalated at e = 0.3% per month for postponed starts, a risk premium weighted by λ on the predicted P90–P50 spread, and a delay cost of δ = 4% of contract value per month of expected lateness — corresponding to a contractual penalty of about 1‰ per day plus extended site overheads — capped at κ = 15%. Lateness Θ in Eq. (11) is evaluated on a buffered duration at confidence level α, using the historical log-slippage distribution with mean μ = 0.03 and a standard deviation σ_{pm} = 0.07 + 0.25·max(0, 1 − r_{pm}) that grows with schedule compression ratio r. Constraint (12) assigns exactly one mode and start time to each project; (13) limits monthly use of each resource to its capacity R_k; and (14) enforces release dates ϱ_p and a maximum tolerated planned delay of Δ = 2 months beyond due date D_p. The resulting MILP has up to 956 binary variables and 116 constraints and is solved to optimality with the HiGHS solver [39] through SciPy. Capacities in the base case are set at 50% of the aggregate peak demand, which makes the portfolio genuinely resource-constrained.

Three strategies are compared. **S1** represents current practice: all projects in normal mode, scheduled by a serial schedule-generation scheme in earliest-due-date order at the earliest resource-feasible start. **S2** solves the MILP with the ML point estimates only (λ = 0, no duration buffer). **S3** is the proposed risk-aware MILP (λ = 0.5, α = 0.8). Each schedule is evaluated by 5,000 Monte-Carlo realisations of the true cost-generating process, including cost noise and duration slippage, and summarised by expected cost, P90 cost, CVaR at 90% [32], the expected number of late projects and expected delay-months.

Implementation

All analyses were implemented in Python 3.13 with scikit-learn [40], SciPy [41], pandas and Matplotlib, using a fixed random seed for reproducibility. Paired differences in absolute percentage error were tested with the Wilcoxon signed-rank test [42].

RESULTS AND DISCUSSION

Predictive Performance

Table 2 reports cross-validated accuracy on the training period and accuracy on the unseen 2023–2025 test projects. The conventional unit-rate estimate has a test MAPE of 19.1% and estimates only 28% of projects within ±10%, confirming that quantity-based pricing systematically misses the risk-driven component of cost.

Table 2. Accuracy of the cost estimation models (CV: 10-fold cross-validation on 2015–2022; Test: hold-out 2023–2025, n = 138)

Model	CV MAPE (%)	Test MAPE (%)	Test MdAPE (%)	Test RMSE (M EUR)	Test R^2	Within ±10% (%)
Conventional unit-rate estimate	17.4	19.1	16.4	1.398	0.813	28.3
Direct MLR	9.2	9.1	6.9	0.629	0.962	65.9
Direct SVR	9.0	11.6	9.7	1.133	0.877	51.4
Direct Random forest	16.3	18.2	15.5	1.583	0.760	31.2
Direct ANN	17.2	25.2	19.0	3.044	0.112	32.6

Model	CV MAPE (%)	Test MAPE (%)	Test MdAPE (%)	Test RMSE (M EUR)	Test R^2	Within ±10% (%)
Direct Gradient boosting	11.9	11.8	10.2	1.128	0.878	49.3
Hybrid MLR	8.4	8.0	6.2	0.547	0.971	67.4
Hybrid SVR	8.6	9.3	7.1	0.708	0.952	59.4
Hybrid Random forest	9.1	9.3	7.8	0.580	0.968	63.0
Hybrid ANN	12.9	17.0	13.6	1.266	0.847	37.0
Hybrid Gradient boosting	8.1	8.2	6.9	0.525	0.974	71.0

All ML models improve on the conventional estimate, but the formulation matters more than the algorithm. In the direct formulation, random forest and ANN deteriorate sharply on the test set (test MAPE 18.2% and 25.2%), and direct SVR loses 2.7 percentage points relative to its cross-validated error. The reason is structural: the 2023–2025 projects were started at cost-index levels above any value observed during training, and tree ensembles cannot extrapolate beyond the range of the training targets, while the small sample limits the ANN. The hybrid formulation removes this weakness because price level and quantities enter through the engineering offset; every hybrid learner is more accurate than its direct counterpart. Hybrid GBM achieved the lowest cross-validated error and was selected as the reference model; on the test set it attains MAPE = 8.15%, R^2 = 0.974 and estimates 71% of projects within ±10%. Hybrid MLR is statistically indistinguishable on the test set (MAPE = 8.04%; Wilcoxon p = 0.93), while the improvement of hybrid GBM over the conventional estimate is highly significant ($p < 0.001$). The practical lesson for data-scarce environments is that embedding domain knowledge in the model structure delivers more robustness than increasing algorithmic complexity, a finding consistent with tabular-data benchmarks [22,23]. Figure 3 visualises the systematic underestimation of the conventional estimate, the scale-dependent bias of the best direct model and the tight fit of the hybrid GBM.

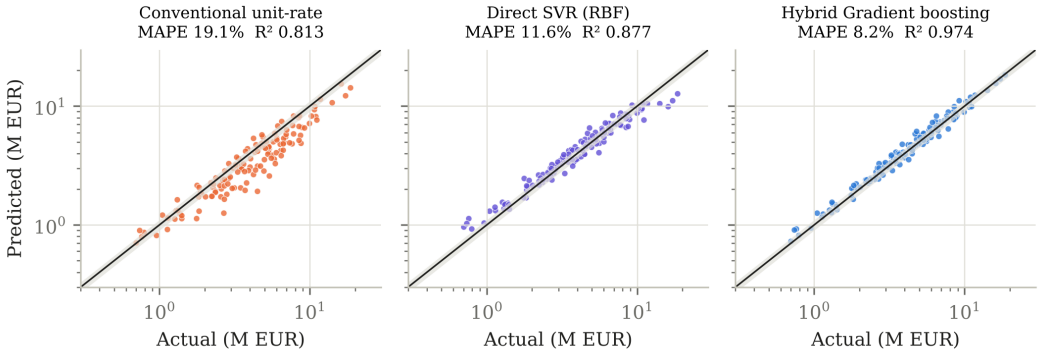


Figure 3. Predicted versus actual cost on the 2023–2025 test projects (log scales; the shaded band marks ±10%)

Drivers of Cost Deviation

Permutation importance of the hybrid GBM (Figure 4a) shows that design completeness at tender explains 76% of the model's predictive importance for cost deviation, followed

by site difficulty (9.5%), region (5.8%), funding source (2.6%), procurement method (1.7%) and the number of bidders (1.6%). Quantity variables carry almost no importance in the hybrid model because their effect is already captured by the unit-rate offset. Partial dependence (Figure 4b) reveals a pronounced non-linearity: cost deviation falls steeply as design completeness rises from 40% to about 75%, and flattens thereafter. Other things equal, a project tendered with a 40% complete design is expected to cost roughly 50% more than one tendered at 80%. Contractor experience (Figure 4c) shows a learning effect of about 5–6% over the first 15 years, while moving from a single bidder to six or more reduces expected cost by about 9%, and the most difficult sites cost about 17% more than the easiest. These results support practical policies such as requiring a minimum design maturity before tendering and promoting competitive procurement, which are directly actionable for Albanian public clients.

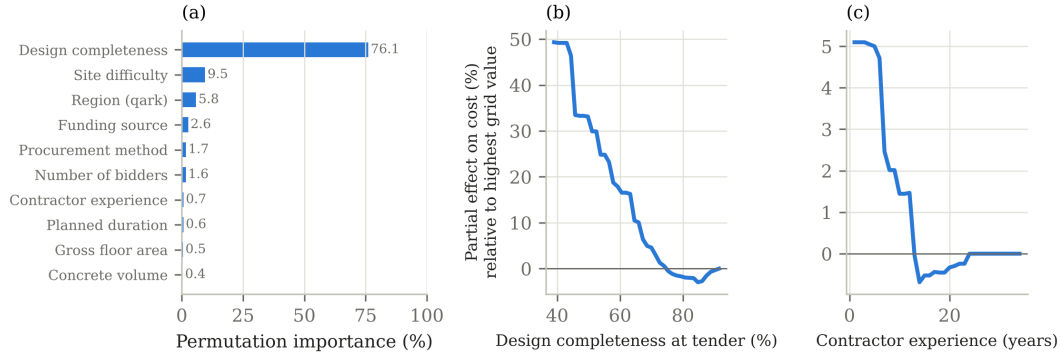


Figure 4. (a) Permutation importance of the hybrid GBM; partial dependence of cost on (b) design completeness and (c) contractor experience

Calibrated Cost Intervals

Uncalibrated quantile GBM intervals covered only 66.7% of test outcomes against a nominal 80%. Conformal calibration raised coverage to 75.4% with an average relative width of 27.3% of the point estimate (Figure 5). The remaining shortfall is explained by the temporal drift between calibration (2021–2022) and test (2023–2025) periods, which violates the exchangeability assumption on which conformal guarantees rest [26]; rolling recalibration as new projects are completed is therefore recommended. Even so, the intervals identify the projects with the widest risk — typically those with immature designs, single bids and difficult sites — and provide a principled basis for contingency allowances.

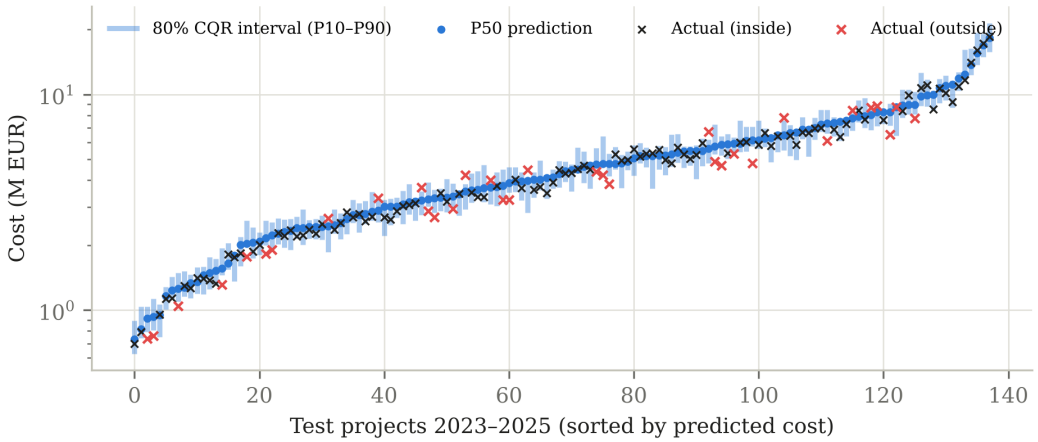


Figure 5. Conformalised 80% prediction intervals and actual costs for the 138 test projects, sorted by predicted cost

Resource Allocation Results

The 14-project portfolio is summarised in Table 3. Its conventional estimate totals EUR 54.9 million, whereas the ML P50 for normal-mode execution totals EUR 75.5 million, highlighting the budgeting gap that the ML stage exposes before any scheduling decision is made.

Table 3. Project portfolio used in the allocation study (Manuf. = manufacturing plant, Wareh. = warehouse and logistics, WWTP = water and wastewater treatment plant; release, due date and durations in months for accelerated/normal/extended modes; costs in M EUR for normal mode)

ID	Type	Region	Area (m²)	Design (%)	Release	Due	Duration (months)	Conv. est.	P10–P90	P50
P1	Agro-food	Fier	3,920	57	0	11	6/8/10	2.81	3.22–4.37	3.97
P2	Energy	Vlorë	4,540	89	0	11	7/9/11	2.79	2.48–3.27	2.84
P3	Manuf.	Elbasan	11,140	59	1	17	10/13/16	8.24	9.40–12.13	10.88
P4	Manuf.	Durrës	6,730	49	1	16	9/11/14	4.94	6.29–7.67	7.38
P5	Wareh.	Fier	2,380	80	1	10	6/7/9	1.13	1.14–1.47	1.33
P6	Manuf.	Tiranë	6,920	66	1	16	9/12/15	4.85	5.69–6.94	6.24
P7	Mining	Shkodër	5,120	84	1	13	8/10/13	4.61	4.39–5.64	4.70
P8	Manuf.	Tiranë	5,860	90	2	16	8/10/13	4.42	4.68–5.73	5.46
P9	Energy	Korçë	2,880	45	2	12	6/7/9	2.01	2.50–3.21	2.79
P10	Manuf.	Dibër	6,830	52	2	17	9/12/15	4.88	6.45–8.76	8.76
P11	WWTP	Tiranë	3,380	83	3	13	7/8/10	3.16	3.28–4.11	4.11
P12	WWTP	Durrës	5,190	61	6	17	8/10/13	4.27	5.13–7.07	6.81
P13	Manuf.	Dibër	4,070	35	6	16	7/8/11	2.84	3.68–5.25	4.79
P14	Agro-food	Dibër	5,170	72	7	21	8/10/13	3.94	4.06–5.81	5.46

Table 4 compares the three strategies at the base-case capacity, and Figure 6 shows the schedules and resource utilisation profiles of S1 and S3.

Table 4. Monte-Carlo evaluation of allocation strategies (5,000 realisations; capacity = 50% of aggregate peak demand)

Indicator	S1 Conventional EDD	S2 ML deterministic	S3 ML risk-aware
Expected cost (M EUR)	76.60	74.33	74.20
P90 cost (M EUR)	79.11	76.80	76.58
$CVaR_{90}$ (M EUR)	80.17	77.89	77.62
Standard deviation (M EUR)	1.96	1.93	1.87
Expected late projects	3.86	3.85	2.87
Expected delay (project-months)	14.5	5.5	4.8
Makespan (months)	22	21	21
Modes (acc./norm./ext.)	0/14/0	3/3/8	3/6/5
Max. planned delay (months)	4	1	2
Mean utilisation, skilled crews (%)	65.8	67.9	70.0
Change in expected cost vs. S1 (%)	–	-2.95	-3.13

The risk-aware MILP (S3) reduces expected portfolio cost from EUR 76.60 million to EUR 74.20 million (3.1%, ≈ EUR 2.4 million), the P90 cost by 3.2% and $CVaR_{90}$ by 3.2%. It also improves delivery: the expected number of late projects falls from 3.86 to 2.87 (26%) and expected delay-months from 14.5 to 4.8 (67%), with a makespan one month shorter. The mechanism is visible in Figure 6. Instead of running all projects at normal pace and pushing late-released projects beyond their due dates, S3 accelerates short projects with tight deadlines, extends projects with slack, and staggers starts so that resource utilisation is high and flat (close to 100% between months 3 and 13) rather than peaking and then idling. The deterministic variant S2 achieves a similar mean cost but leaves 3.85 projects late on average because it schedules completions exactly at due dates and ignores duration risk; the duration buffer and risk premium in S3 convert that hidden risk into an explicit cost, which is precisely the value of propagating uncertainty into the decision.

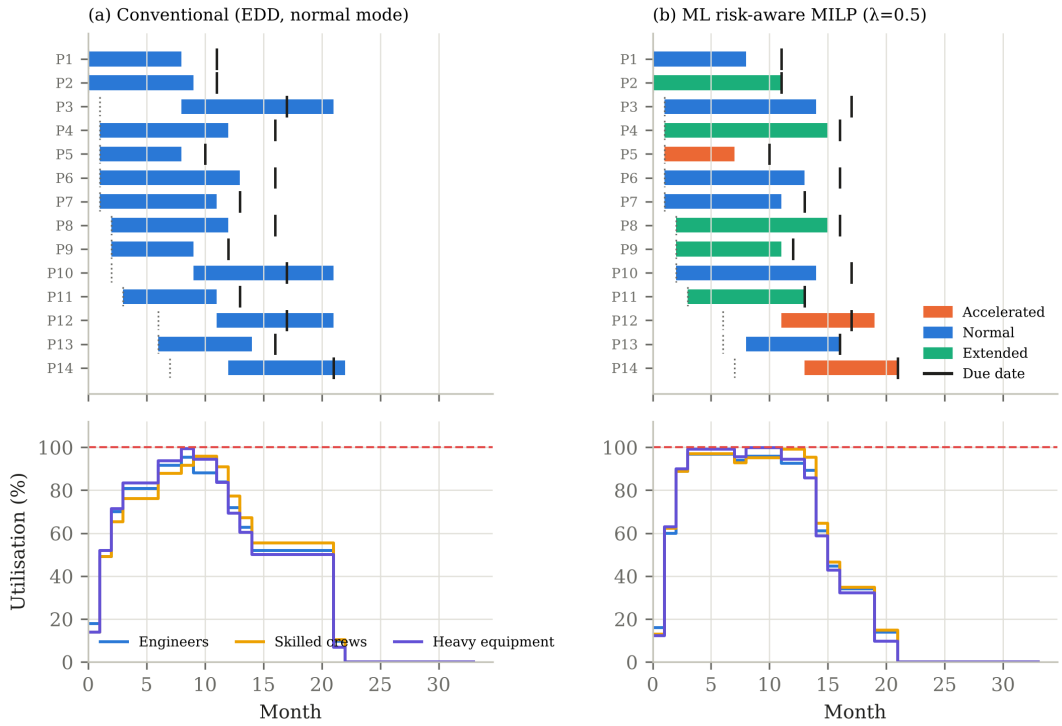


Figure 6. Schedules (top) and monthly resource utilisation (bottom) under (a) conventional earliest-due-date practice and (b) the ML risk-aware MILP; vertical black marks are due dates and dotted marks release dates

Sensitivity Analysis

Figure 7(a) and Table 5 show how the value of optimization changes with resource availability. The benefit is largest when capacity is scarce: at 50–60% of aggregate peak demand, S3 reduces expected cost by 2.0–3.1% and the expected number of late projects by 26–71%. At 45%, the tightest capacity that still admits a schedule within the delay limit, S3 remains 1.2% cheaper but spreads short delays over more projects (5.5 versus 3.8 late projects). Below 45%, no schedule can respect the two-month delay limit; the conventional rule still produces a schedule, but only by accepting much longer delays. When capacity is ample (≥ 70%), all strategies converge, and S1 becomes marginally cheaper (by about 0.3%) because residual ML estimation error slightly distorts the mode choices; in other words, optimization pays off precisely when resources bind, which is the typical situation of regional contractors during investment peaks.

Table 5. Sensitivity of expected portfolio cost (M EUR) and expected late projects to resource capacity

Capacity (% of peak)	S1 cost	S1 late	S3 cost	S3 late	S3 vs. S1 (%)
35	79.70	6.9	infeasible	–	–
40	77.85	5.9	infeasible	–	–
45	76.72	3.8	75.82	5.5	-1.17
50	76.58	3.9	74.20	2.9	-3.11
60	74.20	2.4	72.73	0.7	-1.98

Capacity (% of peak)	S1 cost	S1 late	S3 cost	S3 late	S3 vs. S1 (%)
70	72.17	0.1	72.40	0.2	0.33
100	72.15	0.1	72.40	0.2	0.35

Figure 7(b) varies the duration-buffer confidence level α. Raising α from 0.5 to 0.8 reduces the expected number of late projects from about 3.3 to 2.9 at a negligible cost increase (< 0.2%); beyond 0.8 the solution is unchanged. The risk-premium weight λ had a smaller effect because the predicted P90–P50 spreads vary little across modes of the same project; α is therefore the more effective lever for schedule reliability.

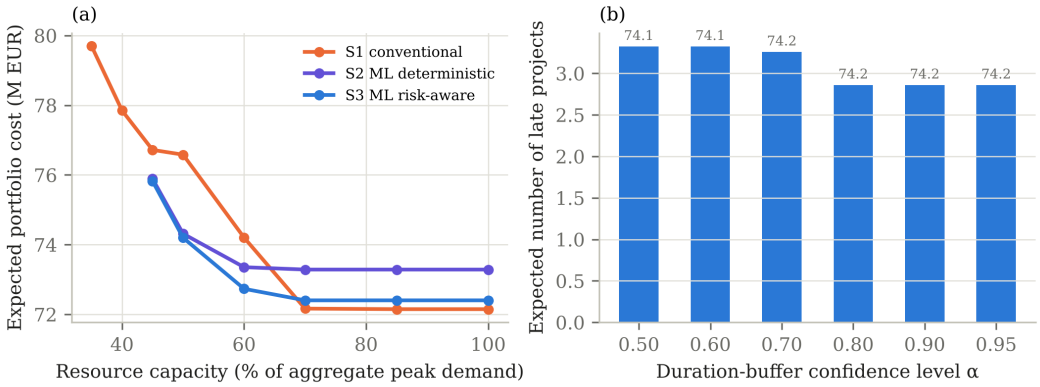


Figure 7. (a) Expected portfolio cost versus resource capacity; (b) expected number of late projects versus duration-buffer confidence level α (labels give expected cost, M EUR)

Managerial Implications and Limitations

For contractors, the framework turns historical project records into calibrated budgets and risk-aware schedules that can be re-optimised as capacities or priorities change. For public clients and the procurement system, the dominant influence of design completeness and competition indicates that the largest savings lie before contract award: tendering on mature designs and attracting several bidders reduces both the level and the variability of cost. For policy-makers, the instability of direct ML models under the 2022–2023 price shock shows that any data-driven estimating tool should be anchored to an official cost index, such as INSTAT's CCI, and recalibrated regularly.

The study has limitations. The case study is a calibrated simulation rather than a census of observed Albanian projects, so the reported magnitudes should be read as a demonstration of the method rather than as estimates of national overrun rates; validation on proprietary contractor databases is the essential next step. The resource model uses uniform monthly demand and three aggregated resources; activity-level precedence networks, multi-skilled labour and stochastic resource availability would increase realism. Finally, the conformal intervals assume exchangeability, which is only approximately satisfied under price drift; adaptive conformal methods could address this.

CONCLUSION

This paper proposed and demonstrated an integrated machine learning and optimization framework for cost estimation and resource allocation in industrial engineering projects, using a calibrated Albanian case study of 420 projects and a 14-project portfolio. Three conclusions emerge. First, hybrid models that learn the deviation from an index-adjusted unit-rate estimate are both accurate and robust: the hybrid gradient boosting model reduced the out-of-time MAPE from 19.1% to 8.2%, whereas direct tree-based models failed under post-2021 escalation. Second, design completeness at tender, site difficulty, region and competition are the dominant drivers of cost deviation, with strongly non-linear effects that conventional estimating ignores. Third, propagating calibrated cost and duration uncertainty into a mixed-integer resource allocation model lowered expected portfolio cost by 3.1% and expected delay by 67% relative to current scheduling practice, with the largest gains when resources are scarce.

The advantages of the framework are its transparency, modest data requirements and reproducibility with open-source tools; its main limitation is the need for validation on observed project records. Possible applications include tender pricing and contingency setting by contractors, portfolio planning by industrial investors, and ex-ante appraisal by public agencies. Future work will apply the framework to observed Albanian contractor and procurement data, extend the optimization to activity-level networks with stochastic programming, and explore online recalibration of the prediction intervals.

ACKNOWLEDGMENT

The author thanks colleagues of the Department of Economic Engineering at the Polytechnic University of Tirana for helpful discussions.

CONFLICT OF INTERESTS

The author confirms that there is no conflict of interest associated with this publication.

DATA AVAILABILITY

The calibrated project database (albania_industrial_projects.csv) and the complete Python code that generates the data, trains the models, solves the optimization model and reproduces all tables and figures are available from the corresponding author on reasonable request.

REFERENCES

1. Flyvbjerg B., Holm M.S. and Buhl S. Underestimating Costs in Public Works Projects: Error or Lie?, *Journal of the American Planning Association*, **2002**; 68(3); pp. 279-295.
2. Odeck J. Cost Overruns in Road Construction — What Are Their Sizes and Determinants?, *Transport Policy*, **2004**; 11(1); pp. 43-53.
3. Love P.E.D. and Ahiaga-Dagbui D.D. 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*, **2018**; 113; pp. 357-368.

4. Ahiaga-Dagbui D.D. and Smith S.D. Rethinking Construction Cost Overruns: Cognition, Learning and Estimation, *Journal of Financial Management of Property and Construction*, **2014**; 19(1); pp. 38-54.
5. INSTAT. Construction Cost Index for Dwellings, Q2-2025. Institute of Statistics of Albania, Tirana, **2025**. Available online: https://www.instat.gov.al/en/themes/prices/construction-cost-index/publications/2025/construction-cost-index-for-dwellings-q2-2025/ (accessed on 8 November 2026).
6. INSTAT. Wage Statistics, Q1-2025. Institute of Statistics of Albania, Tirana, **2025**. Available online: https://www.instat.gov.al/media/4bncjxqf/wage-statistics-q1-2025.pdf (accessed on 8 November 2026).
7. Elmousalami H.H. Artificial Intelligence and Parametric Construction Cost Estimate Modeling: State-of-the-Art Review, *Journal of Construction Engineering and Management*, **2020**; 146(1); 03119008.
8. Tayefeh Hashemi S., Ebadati O.M. and Kaur H. Cost Estimation and Prediction in Construction Projects: A Systematic Review on Machine Learning Techniques, *SN Applied Sciences*, **2020**; 2; 1703.
9. Hegazy T. Optimization of Resource Allocation and Leveling Using Genetic Algorithms, *Journal of Construction Engineering and Management*, **1999**; 125(3); pp. 167-175.
10. Kolisch R. and Hartmann S. Experimental Investigation of Heuristics for Resource-Constrained Project Scheduling: An Update, *European Journal of Operational Research*, **2006**; 174(1); pp. 23-37.
11. Bertsimas D. and Kallus N. From Predictive to Prescriptive Analytics, *Management Science*, **2020**; 66(3); pp. 1025-1044.
12. Elmachtoub A.N. and Grigas P. Smart "Predict, then Optimize", *Management Science*, **2022**; 68(1); pp. 9-26.
13. Hegazy T. and Ayed A. Neural Network Model for Parametric Cost Estimation of Highway Projects, *Journal of Construction Engineering and Management*, **1998**; 124(3); pp. 210-218.
14. Emsley M.W., Lowe D.J., Duff A.R., Harding A. and Hickson A. Data Modelling and the Application of a Neural Network Approach to the Prediction of Total Construction Costs, *Construction Management and Economics*, **2002**; 20(6); pp. 465-472.
15. Kim G.H., An S.H. and Kang K.I. Comparison of Construction Cost Estimating Models Based on Regression Analysis, Neural Networks, and Case-Based Reasoning, *Building and Environment*, **2004**; 39(10); pp. 1235-1242.
16. Cheng M.Y., Tsai H.C. and Sudjono E. Conceptual Cost Estimates Using Evolutionary Fuzzy Hybrid Neural Network for Projects in Construction Industry, *Expert Systems with Applications*, **2010**; 37(6); pp. 4224-4231.
17. Matel E., Vahdatikhaki F., Hosseinyalamdary S., Evers T. and Voordijk H. An Artificial Neural Network Approach for Cost Estimation of Engineering Services, *International Journal of Construction Management*, **2022**; 22(7); pp. 1274-1287.
18. Smola A.J. and Schölkopf B. A Tutorial on Support Vector Regression, *Statistics and Computing*, **2004**; 14(3); pp. 199-222.
19. Breiman L. Random Forests, *Machine Learning*, **2001**; 45(1); pp. 5-32.
20. Friedman J.H. Greedy Function Approximation: A Gradient Boosting Machine, *Annals of Statistics*, **2001**; 29(5); pp. 1189-1232.
21. Chen T. and Guestrin C. XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA, **2016**; pp. 785-794.

22. Grinsztajn L., Oyallon E. and Varoquaux G. Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data?, *Advances in Neural Information Processing Systems 35 (Datasets and Benchmarks Track)*, **2022**.
23. Shwartz-Ziv R. and Armon A. Tabular Data: Deep Learning Is Not All You Need, *Information Fusion*, **2022**; 81; pp. 84-90.
24. Koenker R. and Bassett G. Regression Quantiles, *Econometrica*, **1978**; 46(1); pp. 33-50.
25. Meinshausen N. Quantile Regression Forests, *Journal of Machine Learning Research*, **2006**; 7; pp. 983-999.
26. Vovk V., Gammerman A. and Shafer G. *Algorithmic Learning in a Random World*; Springer: New York, USA, **2005**.
27. Romano Y., Patterson E. and Candès E. Conformalized Quantile Regression, *Advances in Neural Information Processing Systems 32*, **2019**; pp. 3543-3553.
28. Pritsker A.A.B., Watters L.J. and Wolfe P.M. Multiproject Scheduling with Limited Resources: A Zero-One Programming Approach, *Management Science*, **1969**; 16(1); pp. 93-108.
29. Blazewicz J., Lenstra J.K. and Rinnooy Kan A.H.G. Scheduling Subject to Resource Constraints: Classification and Complexity, *Discrete Applied Mathematics*, **1983**; 5(1); pp. 11-24.
30. Feng C.W., Liu L. and Burns S.A. Using Genetic Algorithms to Solve Construction Time-Cost Trade-Off Problems, *Journal of Computing in Civil Engineering*, **1997**; 11(3); pp. 184-189.
31. Herroelen W. and Leus R. Project Scheduling under Uncertainty: Survey and Research Potentials, *European Journal of Operational Research*, **2005**; 165(2); pp. 289-306.
32. Rockafellar R.T. and Uryasev S. Optimization of Conditional Value-at-Risk, *Journal of Risk*, **2000**; 2(3); pp. 21-41.
33. Public Procurement Agency of Albania (Agjencia e Prokurimit Publik). Official Website. Available online: https://www.app.gov.al (accessed on 8 November 2026).
34. Hastie T., Tibshirani R. and Friedman J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, USA, **2009**.
35. Bergmeir C. and Benítez J.M. On the Use of Cross-Validation for Time Series Predictor Evaluation, *Information Sciences*, **2012**; 191; pp. 192-213.
36. Fisher A., Rudin C. and Dominici F. All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously, *Journal of Machine Learning Research*, **2019**; 20(177); pp. 1-81.
37. Lundberg S.M. and Lee S.I. A Unified Approach to Interpreting Model Predictions, *Advances in Neural Information Processing Systems 30*, **2017**; pp. 4765-4774.
38. Chernozhukov V., Fernández-Val I. and Galichon A. Quantile and Probability Curves Without Crossing, *Econometrica*, **2010**; 78(3); pp. 1093-1125.
39. Huangfu Q. and Hall J.A.J. Parallelizing the Dual Revised Simplex Method, *Mathematical Programming Computation*, **2018**; 10(1); pp. 119-142.
40. Pedregosa F., Varoquaux G., Gramfort A., et al. Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, **2011**; 12; pp. 2825-2830.
41. Virtanen P., Gommers R., Oliphant T.E., et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, *Nature Methods*, **2020**; 17; pp. 261-272.
42. Wilcoxon F. Individual Comparisons by Ranking Methods, *Biometrics Bulletin*, **1945**; 1(6); pp. 80-83.