CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (2) (2026), Pages 2678–2688
p-ISSN: 2086-0382; e-ISSN: 2477-3344

# Evaluating Predictive Mathematical Models of Shipboard Operational Decision-Making Performance Using Scenario-Based Assessment

Novitasari, Achmad Dhany Fachrudin, Eka Nurmala Sari Agustina*, I'ie Suwondo, Antonius Edy Kristiyono, and Arif Fadhlur Rohman Fattah

*Politeknik Pelayaran Surabaya, Surabaya, Indonesia*

**Abstract**

Shipboard operational decision making requires seafarers to interpret numerical information, identify operational risks, apply procedures, use onboard technology, and coordinate actions under changing conditions. This study evaluated predictive mathematical models of shipboard operational decision-making performance using scenario-based assessment. Data were obtained from 78 experienced seafarers enrolled in Recognition of Prior Learning programs in Nautical Studies and Marine Engineering. Five self-reported constructs were examined: operational data understanding, seafaring experience, SOP and checklist use, shipboard technology utilization, and communication and coordination. The outcome was the mean of four discipline-specific scenario scores. During manuscript revision, all scenario responses were re-evaluated with the predefined 1–5 rubric by discipline-specific subject-matter lecturers, and all analyses were repeated using these expert-assessed scores. OLS with HC3 robust standard errors was used as the inferential baseline, while Ridge regression was assessed using repeated nested stratified five-fold cross-validation. The OLS model yielded $R^2 = 0.198$ and adjusted $R^2 = 0.130$; $X_1$–$X_5$ were not jointly significant after program adjustment ($p = 0.231$). The best predictive model, Ridge with $X_1$–$X_5$ and program, achieved mean cross-validated $R^2 = 0.075$ (SD = 0.030), RMSE = 1.099, and MAE = 0.950. These findings indicate modest out-of-sample prediction and support cautious use of self-reported operational constructs as predictors of scenario-demonstrated performance.

**Keywords:** decision-making performance; maritime operations; predictive modeling; Ridge regression; scenario-based assessment

Copyright © 2026 by Authors, Published by CAUCHY Group. This is an open access article under the CC BY-SA License (https://creativecommons.org/licenses/by-sa/4.0)

## 1. Introduction

Shipboard operations are performed in a safety-critical environment in which decisions frequently have to be made from multiple sources of operational information under time, environmental, and technical constraints. Modern seafarers are required not only to possess technical knowledge but also to interpret equipment information, recognize abnormal conditions, evaluate alternatives, communicate effectively, and select actions that maintain safe vessel operation. As maritime systems become increasingly digital and interconnected, competence therefore needs to be understood as an interaction between human capabilities and shipboard technological systems rather than as technical knowledge alone [1].

*Corresponding author. E-mail: eka.nurmala@poltekpel-sby.ac.id

Received: August 23, 2026 Revised: October 29, 2026 Accepted: November 7, 2026
DOI: https://doi.org/10.18860/cauchy.v11i2.46178

Performance-based and scenario-oriented assessment has consequently become increasingly important in maritime education and training. A review of simulator-based maritime education identifies authentic assessment, human factors, performance measurement, and competence evaluation as recurring themes in the field [2]. Recent work also shows that simulator assessment can become overly dependent on instructor judgment unless performance criteria and scoring procedures are made explicit [3], while automatic assessment studies demonstrate that carefully programmed scenarios and evaluation criteria can support more objective competence measurement [4].

Several dimensions may theoretically contribute to operational decision making. Understanding numerical and operational data is necessary for interpreting changes in speed, distance, temperature, pressure, engine performance, traffic information, and other shipboard indicators. Seafaring experience, procedural compliance, shipboard technology use, and communication are also plausible predictors, and experimental maritime research has shown that higher-performing teams display distinctive communication structures and richer situation-aware communication content [5].

Recent data-driven maritime research further demonstrates that decision making, situational awareness, communication, procedures, information processing, and maneuvering competence are closely connected to maritime safety outcomes [6]. At the same time, objective human-performance research using maritime simulation indicates that demonstrated performance may require direct performance measures rather than relying only on subjective judgments or self-reported competence [7]. Adaptive human-factor models likewise emphasize the interaction among cognitive behavior, operational context, and multiple risk influences in maritime decision processes [8].

The design of the scenarios themselves is also important because scenario-based assessment should correspond functionally to real shipboard practice while remaining aligned with clear assessment objectives. Recent maritime simulator research argues that effective scenarios require functional congruence, suitable complexity, and alignment with the participants' background and intended outcomes [9]. This perspective supports the use of discipline-specific operational cases in which Nautical and Marine Engineering respondents are assessed using scenarios relevant to their professional domains.

A second methodological issue concerns the distinction between explaining an observed dataset and predicting new observations. Multiple linear regression may provide coefficient estimates and inferential tests, but correlated predictors can destabilize individual coefficients and produce optimistic interpretations if model quality is judged only from in-sample $R^2$. Ridge regression provides a regularized alternative by shrinking regression coefficients and is widely used when predictors are strongly correlated [10–12]. Predictive evaluation also requires strict separation between model tuning and performance estimation; nested and repeated cross-validation are established approaches for reducing optimistic model-selection bias and for obtaining more defensible estimates of out-of-sample performance [13–18].

Accordingly, this study aims to evaluate predictive mathematical models of shipboard operational decision-making performance using scenario-based assessment. Five self-reported operational constructs are examined as predictors: operational data understanding, seafaring experience, SOP and checklist use, shipboard technology utilization, and communication and coordination. OLS is used as an inferential baseline, while Ridge regression is evaluated for out-of-sample prediction under correlated predictors. The principal contribution is to quantify predictive generalization rather than relying only on in-sample statistical significance and to examine the extent to which self-reported operational factors add predictive information beyond discipline/scenario-set context.

# 2. Methods

This section presents the methodological framework employed in this study, covering the research design, participant characteristics, predictor variables, and scenario-based performance assessment. It also describes the statistical modeling procedures, predictive validation strategy, and evaluation metrics used to assess model performance.

## 2.1. Research design and participants

This study employed a quantitative cross-sectional predictive modeling design. Data were collected through an online questionnaire from experienced seafarers enrolled in Recognition of Prior Learning (*Rekognisi Pembelajaran Lampau*, RPL) programs in Nautical Studies and Marine Engineering at Politeknik Pelayaran Surabaya. Respondents were instructed to answer according to their actual onboard experience. Data collection was conducted between 7 July and 27 August 2026. A total of 83 questionnaire submissions were obtained. Duplicate submissions were identified using respondent names, and when a respondent submitted more than once, the chronologically latest response was retained. This procedure resulted in a final analytical sample of $n = 78$, comprising 37 RPL Nautika and 41 RPL Teknika respondents. Missingness was assessed after duplicate handling. No missing values were identified in the 20 predictor items, the program indicator, or the four discipline-specific scenario responses applicable to each participant. Blank cells corresponding to the other discipline's scenario set were structurally not applicable rather than missing; consequently, no imputation was performed.

The sample represented experienced maritime personnel. Of the 78 respondents, 52 (66.7%) reported more than 10 years of seafaring experience, 16 (20.5%) reported 7–10 years, and 10 (12.8%) reported 4–6 years. In addition, 59 respondents (75.6%) indicated that they had previously encountered abnormal operational conditions onboard ships. Professional ranks were standardized during data preparation so that semantically equivalent entries such as KKM and Chief Engineer, or Mualim I and Chief Officer, were assigned to common categories while original responses were retained in the raw dataset.

## 2.2. Predictor variables

Each of the five predictor constructs was measured using four statements on a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). $X_1$ represented Operational Data Understanding; $X_2$ Seafaring Experience; $X_3$ SOP and Checklist Use; $X_4$ Shipboard Technology Utilization; and $X_5$ Communication and Coordination. For each construct, the arithmetic mean of its four items was used as the predictor score. The complete questionnaire is provided in the Supplementary Material.

The five predictor constructs showed high internal consistency, with Cronbach's alpha coefficients ranging from 0.931 to 0.986. However, coefficients of this magnitude may also reflect substantial item overlap or redundancy. These alpha coefficients were therefore interpreted as measures of internal consistency rather than evidence of content validity, construct validity, or empirical separation among the five constructs. Formal expert-based content validation was not conducted prior to data collection.

## 2.3. Scenario-based decision-making performance

The dependent variable $Y$ represented Scenario-Based Operational Decision-Making Performance. Each respondent completed four operational scenarios appropriate to the respondent's discipline. Nautical respondents evaluated estimated time of arrival under reduced speed, collision risk based on CPA/TCPA information, route deviation caused by wind and current, and vessel delay under combined operational constraints. Marine Engineering respondents evaluated main-engine overheating, lubricating-oil pressure reduction, inefficient fuel-consumption–RPM behavior, and generator overload or failure situations.

Responses were evaluated using a predefined analytical rubric ranging from 1 to 5. A score of 1 represented a response with limited relevance and no clear operational reasoning, whereas a score of 5 represented a response that appropriately interpreted the available operational information, considered safety and/or efficiency, and proposed a logical operational action. Intermediate scores represented progressively stronger integration of data interpretation, causal reasoning, procedural considerations, and operational decision making.

To strengthen the outcome assessment during manuscript revision, all scenario responses were re-evaluated using the same predefined rubric by discipline-specific subject-matter lecturers. Nautical responses were assessed by one lecturer in Nautical Studies, whereas Marine Engineering responses were assessed by one lecturer in Marine Engineering. Respondent identities were represented by coded identifiers, and the raters were not provided with the previous provisional scenario scores or the $X_1$–$X_5$ predictor scores during the re-evaluation. All analyses reported in the revised manuscript were repeated using these expert-assessed scenario scores.

$$Y_i = \frac{S_{i1} + S_{i2} + S_{i3} + S_{i4}}{4}, \tag{1}$$

where $S_{ik}$ denotes the expert-assessed score for respondent $i$ on scenario $k$. As shown in Eq. (1), the revised outcome was the mean of the four scenario scores. The outcome ranged from 1 to 5 and showed high internal consistency across the four scenario scores (Cronbach's $\alpha = 0.944$). Because each disciplinary response set was evaluated by a single discipline-specific rater, independent duplicate ratings were not available and inter-rater reliability could not be estimated. Accordingly, Cronbach's alpha should not be interpreted as evidence of inter-rater agreement.

## 2.4. Ordinary least squares model

OLS regression was used as the inferential baseline. Program affiliation was included as an adjustment variable because Nautical and Marine Engineering participants completed different discipline-specific scenario sets. It was coded as $P_i = 0$ for RPL Nautika and $P_i = 1$ for RPL Teknika. Because program membership and scenario-set content were not separable in the present design, the program indicator reflects both disciplinary context and differences in assessment scenarios.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 P_i + \varepsilon_i. \tag{2}$$

The baseline inferential model was specified in Eq. (2). HC3 heteroskedasticity-consistent standard errors were used for statistical inference, and multicollinearity was evaluated using the variance inflation factor (VIF). The program coefficient was treated only as an adjustment for discipline/scenario-set context and not as evidence that one maritime discipline inherently possessed higher decision-making performance than the other. Statistical significance was evaluated at $\alpha = 0.05$.

## 2.5. Ridge regression

Because several predictor constructs represented related aspects of operational competence and were strongly correlated, a regularized Ridge model was evaluated in addition to OLS. Ridge regression minimizes the residual sum of squares together with an $L_2$ penalty [10–12]:

$$\widehat{\boldsymbol{\beta}}_{\text{Ridge}} = \arg\min_{\boldsymbol{\beta}} \left[ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right]. \tag{3}$$

In Eq. (3), $\lambda \geq 0$ controls coefficient shrinkage. Candidate specifications included $X_1$–$X_5$ only, $X_1$–$X_5$ plus program, and a discipline-sensitive specification containing program-by-predictor interactions. Corresponding OLS specifications and simple benchmarks were evaluated for comparison.

### 2.6. Predictive validation and evaluation metrics

Predictive performance was evaluated using a repeated nested cross-validation framework to separate Ridge hyperparameter tuning from out-of-sample performance estimation [13–15]. The outer procedure consisted of stratified five-fold cross-validation repeated 20 times, preserving the approximate Nautical–Marine Engineering composition across folds. Outer random seeds were fixed at 202600–202619.

Within each outer training fold, Ridge hyperparameter tuning was performed using inner stratified five-fold cross-validation. The penalty $\lambda$ was selected from 41 logarithmically spaced values from $10^{-4}$ to $10^{4}$ by the lowest mean squared error across the inner-validation folds. Inner random seeds followed the fixed sequence $303000 + 10r + f$, where $r$ is the repetition index and $f$ the outer-fold index. Standardization parameters were estimated only from the corresponding training data and then applied to validation or test observations.

Within each outer repetition, every participant received exactly one out-of-fold prediction. Cross-validated $R^2$, RMSE, and MAE were calculated from the complete set of 78 out-of-fold predictions for that repetition, yielding 20 estimates of each metric. Primary model performance is reported as mean ± SD across repetitions. For model comparison, the best predictive model was defined as the candidate with the highest mean cross-validated $R^2$ across the 20 outer repetitions; RMSE and MAE were treated as complementary error measures, and the between-repetition SD was used to describe sensitivity to the repeated fold partitions. The reported 95% confidence interval for mean cross-validated $R^2$ was calculated from the 20 repetition-level estimates as $\overline{R^2} \pm t_{0.975,19} s/\sqrt{20}$. Because these repeated cross-validation estimates are generated from repeated partitions of the same dataset rather than independent external samples, the interval is interpreted as uncertainty around the mean repeated-cross-validation performance under this resampling procedure. It is not a prediction interval for individual observations and should not be interpreted as a guarantee of performance in an independent population. Aggregated out-of-fold metrics, obtained by averaging each participant's 20 out-of-fold predictions, were retained as secondary descriptive summaries and were not used as the primary model-selection statistic.

### 2.7. Statistical software

The analysis was implemented in Python 3.13.5 using NumPy 2.3.5, SciPy 1.17.0, statsmodels 0.14.6, and scikit-learn 1.8.0. ChatGPT (OpenAI) was used as an AI-assisted tool to support code development and statistical presentation during manuscript preparation and revision. All analytical code, parameter settings, numerical outputs, and reported results were reviewed and verified by the authors, and all final analytical decisions and scientific interpretations were made by the authors. Predictive interpretation was based primarily on repeated out-of-sample performance rather than statistical significance alone.

## 3. Results and Discussion

After duplicate handling, the analytical dataset contained 78 respondents. The participant profile is summarized in Table 1.

The descriptive statistics in Table 2 show pronounced concentration of the self-reported predictors toward the upper end of the scale. The revised expert-assessed scenario outcome had a lower mean (2.946; SD = 1.151), indicating greater differentiation in demonstrated performance.

The concentration of predictor scores is illustrated in Fig. 1. Such range restriction may reduce between-participant variation available to predictive models. The predictor constructs were moderately to strongly intercorrelated (Table 3), with Pearson correlations ranging from 0.648 to 0.890. The strongest association occurred between $X_3$ and $X_4$ ($r = 0.890$).

Bivariate associations with the revised outcome were positive and modest (Table 4). $X_3$ and $X_4$ had VIF values above 5, confirming a multicollinearity concern relevant to Ridge regularization.

**Table 1:** Participant profile after data cleaning

| Characteristic | Category | $n$ | % |
|---|---|---|---|
| Program | RPL Nautika | 37 | 47.4 |
| Program | RPL Teknika | 41 | 52.6 |
| Age | $>40$ years | 33 | 42.3 |
| Age | 31–35 years | 18 | 23.1 |
| Age | 36–40 years | 14 | 17.9 |
| Age | 25–30 years | 11 | 14.1 |
| Age | $<25$ years | 2 | 2.6 |
| Experience | $>10$ years | 52 | 66.7 |
| Experience | 7–10 years | 16 | 20.5 |
| Experience | 4–6 years | 10 | 12.8 |

| Characteristic | Category | $n$ | % |
|---|---|---|---|
| Abnormal-condition experience | Yes | 59 | 75.6 |
| Abnormal-condition experience | No | 19 | 24.4 |

**Table 2:** Descriptive statistics, internal consistency, and ceiling indicators

| Construct | Description | Mean | SD | $\alpha$ | $\geq 4$ (%) | $=5$ (%) |
|---|---|---|---|---|---|---|
| $X_1$ | Operational Data Understanding | 4.119 | 0.973 | 0.931 | 78.2 | 30.8 |
| $X_2$ | Seafaring Experience | 4.292 | 0.980 | 0.977 | 84.6 | 43.6 |
| $X_3$ | SOP and Checklist Use | 4.407 | 0.914 | 0.982 | 91.0 | 50.0 |
| $X_4$ | Shipboard Technology Utilization | 4.266 | 0.956 | 0.975 | 83.3 | 41.0 |
| $X_5$ | Communication and Coordination | 4.298 | 1.080 | 0.986 | 85.9 | 48.7 |
| $Y$ | Scenario-Based Decision-Making Performance | 2.946 | 1.151 | 0.944 | – | – |

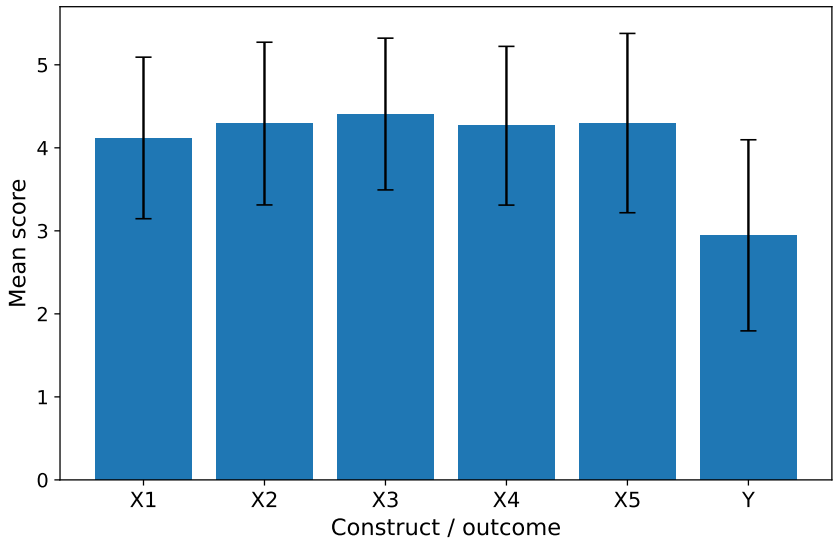



**Fig. 1:** Mean scores of the five predictor constructs and expert-assessed scenario-based performance; error bars represent one standard deviation.

The OLS model with HC3 robust standard errors is reported in Table 5. The model explained 19.8% of the in-sample variance ($R^2 = 0.198$; adjusted $R^2 = 0.130$). None of $X_1$–$X_5$ was individually significant after mutual adjustment, and the five predictors were not jointly significant after program adjustment (HC3 joint $p = 0.231$). The coefficient for program was positive ($\beta = 0.554$, $p = 0.034$), but it cannot be interpreted as an inherent difference in competence because program membership is confounded with discipline-specific scenario content.

Out-of-sample performance provides a stricter test of prediction. The program-only benchmark

**Table 3:** Pearson correlation matrix among predictor constructs

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|
| $X_1$ | 1.000 | 0.784 | 0.742 | 0.730 | 0.648 |
| $X_2$ | 0.784 | 1.000 | 0.747 | 0.717 | 0.665 |
| $X_3$ | 0.742 | 0.747 | 1.000 | 0.890 | 0.707 |
| $X_4$ | 0.730 | 0.717 | 0.890 | 1.000 | 0.657 |
| $X_5$ | 0.648 | 0.665 | 0.707 | 0.657 | 1.000 |

**Table 4:** Association with scenario-based performance and multicollinearity

| Predictor | Pearson $r$ | $p$ | Spearman $\rho$ | $p$ | VIF |
|---|---|---|---|---|---|
| $X_1$ | 0.341 | 0.002 | 0.367 | $< 0.001$ | 3.173 |
| $X_2$ | 0.329 | 0.003 | 0.506 | $< 0.001$ | 3.214 |
| $X_3$ | 0.343 | 0.002 | 0.436 | $< 0.001$ | 5.906 |
| $X_4$ | 0.342 | 0.002 | 0.396 | $< 0.001$ | 5.098 |
| $X_5$ | 0.317 | 0.005 | 0.441 | $< 0.001$ | 2.225 |

**Table 5:** OLS model with HC3 robust standard errors

| Predictor | $\beta$ | Robust SE | 95% CI low | 95% CI high | $p$ |
|---|---|---|---|---|---|
| Intercept | 0.758 | 0.920 | -1.046 | 2.563 | 0.410 |
| $X_1$ | 0.031 | 0.253 | -0.465 | 0.527 | 0.902 |
| $X_2$ | 0.094 | 0.556 | -0.995 | 1.183 | 0.866 |
| $X_3$ | 0.070 | 0.432 | -0.778 | 0.917 | 0.872 |
| $X_4$ | 0.155 | 0.256 | -0.345 | 0.656 | 0.543 |
| $X_5$ | 0.092 | 0.250 | -0.398 | 0.582 | 0.713 |
| Program (Teknika = 1) | 0.554 | 0.261 | 0.043 | 1.065 | 0.034 |

achieved mean CV $R^2 = 0.034$ (SD = 0.025). According to the model-selection rule defined in the Methods, Ridge with $X_1$–$X_5$ and program was the best-performing candidate because it achieved the highest mean CV $R^2$: 0.075 (SD = 0.030; 95% CI 0.061–0.089), with mean RMSE = 1.099 (SD = 0.018) and mean MAE = 0.950 (SD = 0.017). Although the interaction Ridge specification produced a marginally lower mean MAE (0.938), its mean CV $R^2$ was slightly lower (0.073) and its between-repetition variability in $R^2$ was substantially larger (SD = 0.105); it was therefore not designated as the best model. The 95% CI of 0.061–0.089 summarizes uncertainty around the mean repeated-CV $R^2$ across the 20 fold repetitions and does not represent an individual-level prediction interval or an external-validation performance guarantee. The incremental mean CV $R^2$ over the program/scenario-set benchmark was approximately 0.041. The repeated nested cross-validation results for all candidate models are summarized in Table 6.

**Table 6:** Repeated nested cross-validated predictive performance (20 repetitions)

| Model | CV $R^2$ mean (SD) | 95% CI | RMSE mean (SD) | MAE mean (SD) |
|---|---|---|---|---|
| Mean baseline | -0.033 (0.021) | -0.043 to -0.023 | 1.162 (0.012) | 1.011 (0.012) |
| Program only (OLS) | 0.034 (0.025) | 0.022 to 0.045 | 1.124 (0.015) | 0.982 (0.016) |
| $X_1$–$X_5$ (OLS) | -0.109 (0.083) | -0.147 to -0.070 | 1.203 (0.045) | 1.015 (0.030) |
| $X_1$–$X_5$ + Program (OLS) | -0.097 (0.069) | -0.129 to -0.065 | 1.197 (0.037) | 0.990 (0.026) |
| $X_1$–$X_5$ (Ridge) | 0.058 (0.028) | 0.045 to 0.071 | 1.110 (0.016) | 0.965 (0.016) |
| $X_1$–$X_5$ + Program (Ridge) | 0.075 (0.030) | 0.061 to 0.089 | 1.099 (0.018) | 0.950 (0.017) |
| $X_1$–$X_5$ + Program + interactions (OLS) | -0.361 (0.244) | -0.475 to -0.246 | 1.329 (0.118) | 1.035 (0.061) |
| $X_1$–$X_5$ + Program + interactions (Ridge) | 0.073 (0.105) | 0.023 to 0.122 | 1.100 (0.059) | 0.938 (0.036) |

The mean cross-validated $R^2$ values and their between-repetition variability are visualized

in Fig. 2. Ridge models outperformed corresponding unregularized OLS specifications in out-of-sample $R^2$, but this is described as better predictive performance rather than unquantified model "stability."

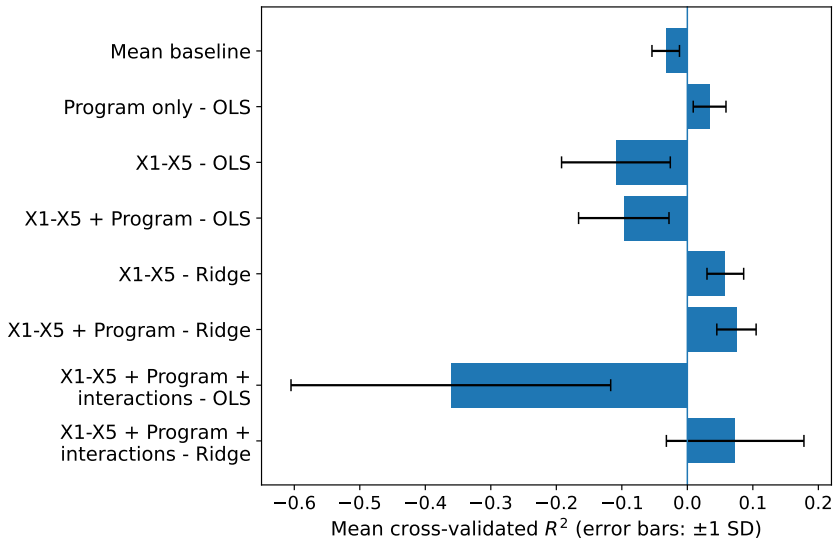



**Fig. 2:** Mean cross-validated $R^2$ for candidate models across 20 repeated outer five-fold cross-validation runs; error bars represent ±1 SD.

These results illustrate the distinction between explanatory and predictive modeling emphasized by Shmueli [19]. The OLS model explained 19.8% of the observed in-sample variance, yet corresponding unregularized predictive models produced negative mean cross-validated $R^2$ values. Ridge improved out-of-sample predictive performance relative to corresponding OLS specifications, but absolute performance remained modest.

The very high internal-consistency coefficients of $X_1$–$X_5$ require cautious interpretation. Although alpha values of 0.931–0.986 indicate strong within-construct consistency, they may also reflect item overlap. This concern is supported by the moderate-to-strong intercorrelations among constructs, particularly $X_3$ and $X_4$ ($r = 0.890$). Pronounced ceiling effects were also present: 91.0% of respondents had an $X_3$ mean of at least 4 and 50.0% obtained the maximum construct mean of 5. Because the predictors were self-reported, socially desirable responding may also have contributed to score concentration. Formal content and discriminant-validity assessments were not conducted; therefore, the extent to which the five constructs represent empirically distinct dimensions cannot be established from the present study.

The program variable requires particular caution. Because Nautical and Marine Engineering participants completed different scenario sets, program membership is statistically inseparable from scenario content. Program-related predictive information may therefore reflect assessment context rather than underlying differences in competence. Future studies should use common anchor scenarios or partially overlapping scenario sets across disciplines.

From an implementation perspective, the present findings should be interpreted as methodological evidence rather than evidence of deployment readiness. The current model should not be used for operational screening, training prioritization, regulatory monitoring, or individual-level decision making. Such applications are potential future directions requiring independent external validation, calibration in new populations, improved measurement, and evaluation against objective operational or simulator-derived outcomes. This cautious position is consistent with previous maritime decision-support work in which interpretable regression models require further operational validation rather than serving as autonomous decision rules [20].

## 4. Conclusion

This study evaluated mathematical models for predicting shipboard operational decision-making performance from five self-reported operational constructs using an expert-assessed scenario-based outcome. The OLS baseline explained a modest proportion of in-sample variance, while $X_1$–$X_5$ did not contribute significantly as a joint block after adjustment for program/scenario-set context. In repeated nested cross-validation, Ridge regression with $X_1$–$X_5$ and program produced the highest mean cross-validated $R^2$, but overall generalization remained modest and the incremental predictive gain over the program-only benchmark was limited.

The findings do not support the self-reported constructs as strong stand-alone predictors of scenario-demonstrated performance. Important measurement limitations remain: the predictor instrument did not undergo formal expert-based content validation before data collection, several constructs displayed very high internal consistency and substantial intercorrelation, and scenario scoring used one discipline-specific rater per domain, preventing estimation of inter-rater reliability. Because program membership was confounded with discipline-specific scenario content, program effects cannot be interpreted as competence differences between Nautical and Marine Engineering participants. Future research should use larger independent samples, formally validated instruments, common anchor scenarios, multiple independent raters, and objective behavioral or simulator-derived measures before practical deployment of predictive models is considered.

## CRediT Authorship Contribution Statement

**Novitasari**: Conceptualization, Methodology, Formal Analysis, Investigation, Data Curation, Writing–Original Draft Preparation, Visualization, Project Administration. **Achmad Dhany Fachrudin**: Methodology, Validation, Supervision, Writing–Review & Editing, Project Administration. **Eka Nurmala Sari Agustina**: Methodology, Validation, Supervision, Writing–Review & Editing, Project Administration, Correspondence. **I'ie Suwondo**: Validation, Resources, Supervision, Writing–Review & Editing. **Antonius Edy Kristiyono**: Validation, Resources, Writing–Review & Editing. **Arif Fadhlur Rohman Fattah**: Investigation, Data Curation, Resources, Writing–Review & Editing. All authors reviewed and approved the final manuscript, the authorship order, the addition of Eka Nurmala Sari Agustina as the third author, and her designation as the corresponding author.

## Declaration of Generative AI and AI-assisted Technologies

During the preparation and revision of this work, the authors used ChatGPT (OpenAI) for language refinement, manuscript structuring, code-development support, statistical presentation, and formatting support. After using this tool, the authors reviewed, verified, and edited the AI-assisted content as necessary. ChatGPT was not used as an author, to generate or fabricate research data, or to make final scientific judgments. The authors take full responsibility for the accuracy, integrity, and final content of the publication.

## Declaration of Competing Interest

The authors declare no competing interests.

## Ethics and Consent to Participate

This study involved a non-interventional online questionnaire administered to adult seafarers. At the time the data were collected, Politeknik Pelayaran Surabaya did not have a dedicated institutional ethics committee or Institutional Review Board (IRB) responsible for the formal ethical review of non-clinical survey-based research of this type. Consequently, no formal

institutional ethics approval or exemption determination/number was obtained prior to data collection. Participation was voluntary, and all respondents provided informed consent through the research questionnaire before participation. Responses were treated confidentially and used solely for academic research purposes. The study involved no medical, clinical, biological, or experimental intervention.

## Funding

This research was self-funded by the authors and received no external funding.

## Acknowledgments

The authors acknowledge Politeknik Pelayaran Surabaya for institutional support and thank the RPL Nautika and RPL Teknika seafarers who participated in this study.

## Data and Code Availability

The anonymized data and analysis code supporting the findings of this study are available from the corresponding author upon reasonable request, subject to participant confidentiality and applicable institutional data-sharing requirements.

## References

[1] S. Fan and Z. Yang. "Analysing seafarer competencies in a dynamic human-machine system". In: *Ocean & Coastal Management* 240 (2023), p. 106662. DOI: 10.1016/j.ocecoaman.2023.106662.

[2] A. C. Wiig, C. Sellberg, and M. Solberg. "Reviewing simulator-based training and assessment in maritime education: A topic modelling approach for tracing conceptual developments". In: *WMU Journal of Maritime Affairs* 22 (2023), pp. 143–164. DOI: 10.1007/s13437-023-00307-4.

[3] M. Karahalil, M. Lützhöft, and J. Scanlan. "Formative assessment in maritime simulator-based higher education". In: *WMU Journal of Maritime Affairs* 22 (2023), pp. 181–207. DOI: 10.1007/s13437-023-00313-6.

[4] W. Juszkiewicz and A. Żukowska. "The use of the K-Sim Polaris simulator in the process of automatic assessment of navigator competence in the aspect of anticollision activities". In: *Applied Sciences* 13.2 (2023), p. 915. DOI: 10.3390/app13020915.

[5] Y. Min, X. Wanyan, S. Liu, X. Wang, Z. Liao, Z.-Q. Tian, C. Liang, and K. Zhu. "Quantitative analysis of team communication for maritime collaborative task performance improvement". In: *International Journal of Industrial Ergonomics* 92 (2022), p. 103362. DOI: 10.1016/j.ergon.2022.103362.

[6] K. Shi, S. Fan, J. Weng, and Z. Yang. "Seafarer competency analysis: Data-driven model in restricted waters using Bayesian networks". In: *Ocean Engineering* 311 (2024), p. 119001. DOI: 10.1016/j.oceaneng.2024.119001.

[7] S. Fan and Z. Yang. "Towards objective human performance measurement for maritime safety: A new psychophysiological data-driven machine learning method". In: *Reliability Engineering & System Safety* 233 (2023), p. 109103. DOI: 10.1016/j.ress.2023.109103.

[8] S. Adumene, M. Afenyo, V. Salehi, and P. William. "An adaptive model for human factors assessment in maritime operations". In: *International Journal of Industrial Ergonomics* 89 (2022), p. 103293. DOI: 10.1016/j.ergon.2022.103293.

[9] S. Harrington, C. Sellberg, and O. Lindwall. "Crafting congruence: Towards systematic scenario design in maritime simulator training and assessment". In: *WMU Journal of Maritime Affairs* 24 (2025), pp. 157–178. DOI: 10.1007/s13437-024-00356-3.

[10] N. Herawati, A. Wijayanti, A. Sutrisno, Nusyirwan, and Misgiyati. "The performance of Ridge Regression, LASSO, and Elastic-Net in controlling multicollinearity: A simulation and application". In: *Journal of Modern Applied Statistical Methods* 23.2 (2024). DOI: 10.56801/Jmasm.V23.i2.4.

[11] I. S. Dar, S. Chand, M. Shabbir, and B. M. G. Kibria. "Condition-index based new ridge regression estimator for linear regression model with multicollinearity". In: *Kuwait Journal of Science* 50.2 (2023), pp. 91–96. DOI: 10.1016/j.kjs.2023.02.013.

[12] M. S. Khan, A. Ali, M. Suhail, E. S. Alotaibi, and N. E. Alsubaie. "On the estimation of ridge penalty in linear regression: Simulation and application". In: *Kuwait Journal of Science* 51.4 (2024), p. 100273. DOI: 10.1016/j.kjs.2024.100273.

[13] M. A. van de Wiel, M. M. van Nee, and A. Rauschenberger. "Fast cross-validation for multi-penalty high-dimensional Ridge Regression". In: *Journal of Computational and Graphical Statistics* 30.4 (2021), pp. 835–847. DOI: 10.1080/10618600.2021.1904962.

[14] S. Varma and R. Simon. "Bias in error estimation when using cross-validation for model selection". In: *BMC Bioinformatics* 7 (2006), p. 91. DOI: 10.1186/1471-2105-7-91.

[15] D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas. "Cross-validation pitfalls when selecting and assessing regression and classification models". In: *Journal of Cheminformatics* 6 (2014), p. 10. DOI: 10.1186/1758-2946-6-10.

[16] S. Bates, T. Hastie, and R. Tibshirani. "Cross-validation: What does it estimate and how well does it do it?" In: *Journal of the American Statistical Association* 119.546 (2024), pp. 1434–1445. DOI: 10.1080/01621459.2023.2197686.

[17] D. Wilimitis and C. G. Walsh. "Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial". In: *JMIR AI* 2 (2023), e49023. DOI: 10.2196/49023.

[18] M. J. Lewis, A. Spiliopoulou, K. Goldmann, C. Pitzalis, P. McKeigue, and M. R. Barnes. "nestedcv: an R package for fast implementation of nested cross-validation with embedded feature selection designed for transcriptomics and high-dimensional data". In: *Bioinformatics Advances* 3.1 (2023), vbad048. DOI: 10.1093/bioadv/vbad048.

[19] G. Shmueli. "To Explain or to Predict?" In: *Statistical Science* 25.3 (2010), pp. 289–310. DOI: 10.1214/10-STS330.

[20] A. D. Fachrudin, N. Novitasari, M. Zainuddin, F. F. Azhar, A. D. Santoso, and N. Basuki. "An MSI-Transformed Multiple Linear Regression Model for Early Warning of Seafarers' Safety Compliance in a Maritime IDSS Prototype". In: *Desimal: Jurnal Matematika* 9.2 (2026), pp. 437–456. DOI: 10.24042/djm.v9i2.32310.