INTERNATIONAL JOURNAL OF ENVIRONMENT, ENGINEERING & EDUCATION

ISSN 2656-8039 (ONLINE)
VOLUME 8 • ISSUE 3 • SEPTEMBER 2026

Received: 2 April 2026 | Accepted: 17 July 2026 | Published: 2 September 2026

Research Article

Machine-Learning Classification of Archived Employability-Readiness Ratings: Concurrent Associations, Calibration, and Held-Out Internal Validation

Tonglu Men • Premsuree Chaumthong *

Department of Education and Society, Rajamangala University of Technology Krungthep, Bangkok, Thailand

ABSTRACT

This study examined the explanatory structure and predictive reproducibility of an end-of-program administrative score used to assess graduate preparedness in vocational education. A cross-sectional secondary analysis was conducted using 450 de-identified learner records collected from four anonymized polytechnic-style institutions during the 2023–2024 academic year. The outcome was a 0–100 composite score analyzed continuously and as three prespecified categories: low (<60), moderate (60–79), and high (≥80). Candidate predictors comprised practical skills, digital competence, soft skills, internship intensity, project performance, and an industry–education integration index treated as an institution-linked contextual proxy. Pearson correlations, heteroskedasticity-consistent ordinary least-squares regression, criterion-overlap sensitivity analysis, and leakage-controlled model development were applied. Six algorithms were compared against a majority-class baseline using a stratified 70:30 development-test split. Bootstrap intervals, ordinal-error measures, and probability-calibration indices were used to quantify uncertainty and performance. The full regression model explained 76% of outcome variance (adjusted R^2 = 0.75), whereas the reduced model retained an adjusted R^2 of 0.58 after removal of predictors susceptible to shared rubric content. In the untouched test partition (n = 135), the back-propagation neural network achieved an accuracy of 0.904 and a macro-F1 of 0.903; all 13 errors occurred between adjacent categories. These findings indicate a coherent within-system scoring structure and strong reproducibility across analytical approaches. However, they do not establish causal effects, transportability across institutions, fairness, operational utility, or prediction of subsequent labor-market outcomes. The model should therefore be restricted to low-stakes auditing, data-quality review, and identification of borderline records pending prospective multi-institutional evaluation.

KEYWORDS algorithmic auditing • criterion contamination • data leakage • educational analytics • industry–education integration • predictive analytics • vocational education

ARTICLE CITATION

T. Men, P. Chaumthong, "Machine-Learning Classification of Archived Employability-Readiness Ratings: Concurrent Associations, Calibration, and Held-Out Internal Validation," International Journal of Environment, Engineering and Education, Vol. 8, No. 3, pp. 360-379, 2026. https://doi.org/10.55151/ijeedu.v8i3.503

***CORRESPONDENCE**

Premsuree Chaumthong premsuree.c@mail.rmutk.ac.th Department of Education and Society, Institute of Science Innovation and Culture, Rajamangala University of Technology Krungthep, 2 Nanglinchee Road, Floor 9, Building 50, Sathorn, Bangkok 10120, Thailand https://orcid.org/0009-0006-9741-6222

1. INTRODUCTION

Vocational education operates within labor markets increasingly shaped by digital transformation, artificial intelligence, changes in occupational structures, and growing demand for technical and interpersonal competencies [1]–[3]. Conditions require programmed evaluation to extend beyond completion rates and certification outcomes. Evaluation must also determine whether graduates possess capabilities that are relevant to occupational performance. Contemporary scholarship conceptualizes employability as a multidimensional configuration of occupational competence, digital capability, transferable skills, career resources, adaptability, and opportunities to apply learning in authentic settings [4]–[7]. From this relational perspective, graduate preparedness is shaped not only by individual attributes but also by curriculum design, workplace exposure, assessment practices, institutional resources, and labor-market conditions.

Conceptual differentiation between employability, work readiness, and employment outcomes is necessary for valid program evaluation. Employability refers to the capabilities, resources, and opportunities that support entry into, participation in, and progression through employment. Work readiness concerns preparation for immediate occupational performance, whereas employment outcomes represent realized results such as job acquisition, retention, earnings, occupational alignment, and employer-evaluated performance. These constructs differ in scope, temporal position, and evidentiary requirements, although they are frequently treated as interchangeable in educational research [8], [9]. Such conflation creates jingle-jangle problems and weakens evaluative and causal interpretations [10], [11]. Consequently, a program-generated readiness score cannot be interpreted as evidence of labor-market success unless it has been validated against independent and preferably prospective employment criteria [12], [13].

Several educational domains may contribute to preparation for occupational performance. Practical competence supports the accurate and safe execution of job-related tasks. Digital competence facilitates information processing, communication, content production, and technology-mediated problem solving. Transferable skills support teamwork, professionalism, communication, and adaptation across occupational contexts [5], [7]. Internships, work-integrated learning, and applied projects may connect these domains by requiring learners to apply disciplinary knowledge under authentic constraints. Their effectiveness, however, depends on task authenticity, placement quality, supervision, feedback, assessment design, and curricular integration rather than duration alone [14], [15]. Ratings produced by placement supervisors or project assessors may therefore represent performance within an educational context rather than subsequent performance in employment. Systematic evidence similarly indicates that employability programs tend to produce more consistent improvements in proximal capabilities and self-perceptions than in independently verified labor-market outcomes, which remain influenced by labor demand, recruitment practices, field of study, and socioeconomic resources [16].

These conceptual distinctions become critical when administrative records are used for programmed monitoring. The archive examined in this study contains a completion rating ranging from 0 to 100, derived from placement-supervisor judgements, capstone-defense information, and an internal readiness rubric. Practical-skills, soft-skills, and project-performance scores were recorded as separate variables within the same administrative system. Separate database fields, however, do not establish construct distinctiveness or measurement independence. Shared rubric content, raters, assessment occasions, and institutional contexts may produce criterion contamination or common-method covariance, thereby increasing observed associations without demonstrating that each indicator independently accounts for readiness [12], [17]. Validity concerns the defensibility of an intended score interpretation rather than the magnitude of a correlation or regression coefficient alone. Sensitivity analysis is therefore required to determine whether the multivariable association pattern remains stable after excluding indicators with the greatest potential for overlap with the administrative outcome.

Machine-learning classification does not resolve deficiencies in target definition or construct validity. A flexible classifier may accurately reproduce low, moderate, and high administrative readiness categories because it learns the structure of the rubric from which those categories were generated. Such internal reproducibility does not constitute evidence of employment forecasting. Predictive estimates may also be inflated when outcome-related information is included among the predictors or when preprocessing, feature selection, and hyperparameter optimization incorporate observations that should remain outside model development [18]. When transportability beyond the development sample is claimed, data partitioning must also account for clustering by institution or program [19]; a defensible analytical design therefore requires preprocessing to be restricted to the development data, candidate models to be compared with interpretable baselines, and final performance to be evaluated in an untouched held-out partition [20]–[23]. Overall accuracy alone is insufficient for evaluating multiclass readiness classification. Assessment should include class-specific discrimination, probability ranking, calibration, uncertainty, and the substantive distance between observed and predicted categories [24], [25]. This requirement is particularly relevant when the outcome is ordinal because an error between adjacent categories differs substantively from an error between the lowest and

highest categories. Explainability methods may support model auditing by indicating how predictor values contribute to individual predictions. Post-hoc explanations, however, cannot correct an invalid target, remove predictor–target contamination, or substitute for comparison with transparent baseline models [26]–[28]. The measurement level of contextual variables imposes a separate inferential constraint. An industry–education integration index may appear to vary at the learner level when a single institutional value is repeated across all learners within the same program. When each institution is represented by only one programmed track, institution, track, and contextual score are structurally confounded. Under these conditions, record-level coefficients cannot identify an independent institutional effect. The contextual index can be analyzed as an institution-linked proxy, but its coefficient cannot be interpreted as evidence that institutional integration causes individual readiness.

Subgroup performance comparisons require equivalent caution. Differences in accuracy, macro-F1, calibration, or ordinal error across programmed groups may reflect subgroup sample size, outcome prevalence, assessment practices, feature distributions, or unequal educational opportunities rather than model fairness. Bias in educational analytics may originate from construct definitions, historical labels, feature availability, optimization criteria, and deployment [29], [30]. Fairness assessment therefore requires clearly identified protected attributes, prespecified fairness criteria, adequate subgroup sample sizes, and analysis of the consequences associated with model-supported decisions. Descriptive differences across programmed groups are insufficient to establish fairness or support fairness certification [31]–[34]. Existing employability analytics research frequently prioritizes predictive accuracy while giving less attention to the interpretation of the prediction target, predictor–target overlap, the measurement level of contextual variables, calibration, ordinal misclassification, and the inferential limits of subgroup comparisons. Generalizability, transparency, and ethical risks have been examined within the broader predictive learning analytics literature [35]. These concerns, however, are rarely integrated with employability construct validity and sensitivity to rubric contamination within a single analytical framework. This omission creates a risk that technically adequate classification performance will be interpreted as evidence of labor-market prediction, causal determinants, institutional effects, model fairness, or operational utility, despite the absence of data capable of supporting those claims.

The present study analyses de-identified administrative records from 450 final-year learners enrolled in engineering, information technology, business services, and health-support programmed during 2023–2024. The analysis addresses four research questions:

1) What pooled bivariate and multivariable associations exist between five learner-level indicators, one institution-linked integration proxy, and the administrative readiness rating?
2) Does the multivariable association pattern remain stable after indicators with the greatest susceptibility to criterion overlap are excluded?
3) How do six candidate classifiers perform in an untouched held-out partition across class-label, probability-ranking, calibration, and ordinal-error criteria?
4) How do the held-out performance metrics of the selected models vary descriptively across the four institution–track groups?

The study evaluates the concurrent structure and internal reproducibility of an archived administrative readiness rating. Its contribution lies in integrating construct interpretation, contamination-sensitive association analysis, leakage-controlled classification, calibration assessment, ordinal-error evaluation, and cautious subgroup comparison within a single analytical design. The scope of inference remains restricted to the available administrative data and the assessment system through which those data were generated. The analysis does not identify causal determinants, verify prediction of subsequent employment, estimate independent institutional effects, certify model fairness, or establish operational utility.

2. LITERATURE REVIEW

2.1. Employability, Work Readiness and Employment Outcomes

Employability is a multidimensional and relational construct rather than a fixed stock of individual skills. Occupational competence, transferable capability, digital competence, adaptability, and career identity acquire value through social and cultural resources, institutional opportunity structures, and labor-market conditions. This position is shared by the psycho-social model of employability, the graduate-capital perspective, and later relational formulations [4]–[6], [36]. In vocational education, employability therefore concerns both what learners can perform and whether educational arrangements enable those capabilities to be transferred into authentic work settings.

Employability, work readiness, and employment outcomes occupy different conceptual and temporal positions. Employability refers to resources and opportunities supporting entry into, participation in, and progression through work. Work readiness denotes preparation for immediate occupational performance, including task, behavioral, and organizational expectations at transition. Employment outcomes are external events such as job acquisition, retention, earnings, occupational alignment, and employer-

evaluated performance [8]–[10]. Treating these constructs as interchangeable creates a jingle-jangle problem that weakens cross-study comparison, score interpretation, and causal inference.

This distinction bounds the interpretation of the archived outcome examined in the present study. The dependent variable is an end-of-program administrative readiness composite derived from placement-supervisor judgement, capstone-defense information, and an internal rubric. It is concurrent with the candidate indicators and does not constitute an independent observation of later employment. Contemporary validity theory locates validity in a proposed interpretation and use, supported by evidence concerning content, response processes, internal structure, relations with other variables, generalizability, and consequences [12], [13], [37]. The score can therefore be interpreted as a programmed-defined indicator of completion-stage readiness, not as verified labor-market employability.

2.2. Vocational Capability Domains

Practical competence is central to vocational education because occupational capability must be demonstrated through performance rather than inferred solely from declarative knowledge. Practice-proximal assessment commonly covers task accuracy, safety, troubleshooting, process handling, completion quality, and adaptation to situational constraints. Large-scale vocational assessment similarly prioritizes authentic tasks, domain-specific reasoning, and transfer into occupational action [38]. Practical-skills ratings should consequently relate to readiness, although the association may be inflated when practical performance also contributes to the final administrative rubric.

Digital competence is now a cross-occupational requirement. DigComp 2.2 defines it through information and data literacy, communication and collaboration, digital-content creation, safety, and problem solving [39]. Digital work also requires cognitive, social, and self-regulatory capability rather than software operation alone [40]. Artificial intelligence, data-intensive production, platform-mediated services, and digitally supported assessment are further reshaping vocational systems [1], [3]. Although recent evidence associates digital employability skills with workforce readiness, conclusions remain dependent on measurement and design [2]. Digital competence is therefore treated as a concurrent learner-level signal, not a demonstrated causal determinant.

Soft skills support communication, teamwork, professionalism, adaptability, self-management, and problem resolution. Reviews consistently position these attributes within employability while reporting differences among student, educator, and employer priorities [7], [41]. Their meaning remains situated: communication and collaboration are enacted through particular tasks, teams, organizational norms, and assessment conditions. Practical, digital, and soft skills should therefore be conceptualized as an integrated capability profile rather than independent additive commodities [4], [5]. Correlations among them may reflect genuine capability convergence, common assessment conditions, or both; regression coefficients cannot be translated directly into intervention effects.

2.3. Experiential and Institutional Readiness Factor

Work-integrated learning connects classroom preparation with authentic tasks, organizational routines, professional relationships, and external feedback. Internships may integrate disciplinary, social, and professional learning. Still, their value depends on task authenticity, supervisory quality, progressive responsibility, feedback, learner agency, assessment alignment, and curricular integration rather than duration alone [14], [15]. Internship intensity measured in months is therefore an exposure indicator, not a direct measure of placement quality. Its association with readiness may also reflect placement selection, employer resources, prior learner advantage, or programmed requirements.

Evidence from employability interventions supports this qualification. Work-integrated learning more consistently improves proximal capabilities and perceived employability than independently verified labor-market outcomes [16]. Digital literacy and career knowledge may also mediate the relationship between workplace learning and perceived employability [42]. Internship exposure is thus theoretically relevant, but additional placement time cannot be assumed to cause improved employment without prospective measures of placement quality, task complexity, supervision, and post-placement outcomes.

Capstone and applied-project performance provide integrative evidence because projects combine disciplinary knowledge, planning, digital tools, communication, collaboration, timeliness, and implementation. Project performance may therefore contribute information beyond isolated competence scores. In the archive, however, capstone-defense information also contributed to the readiness composite, creating criterion proximity. The project coefficient may reflect both genuine transfer of learning and shared measurement, which requires an overlap-sensitive model.

Industry-education integration operates at the contextual level. Employer participation, shared infrastructure, joint curriculum design, workplace mentoring, feedback, and integrated assessment may improve programmed relevance and authenticity. Vocational-modernization research likewise treats sustained institutional collaboration as necessary for aligning curricula, staff capability, equipment, artificial intelligence, and occupational demand [1], [3]. The archived index contains only four institution-level values, each linked to one vocational track. Institution, track, assessment culture, and integration score are therefore structurally confounded. The index can describe

contextual differentiation, but its record-level coefficient cannot identify an independent institutional effect. Such inference would require multiple institutions per track, multiple tracks per institution, and clustered or multilevel analysis.

2.4. Readiness Ratings, Construct Validity and Criterion Overlap

Administrative records are useful for programmed monitoring because they are routinely generated, inexpensive, and connected to educational practice. Their meaning nevertheless depends on the construct map, item content, scoring rules, rater processes, weighting, thresholds, and intended decisions. Separate database fields establish technical separation, not construct distinctiveness. When practical skills, soft skills, project performance, and the final composite share content, raters, occasions, or settings, observed relationships combine trait covariance with method covariance [17], [43], [44].

Criterion contamination occurs when information embedded in a predictor also forms part of the criterion, producing a partly circular association and weakening claims of independent explanation or prediction [12], [45]. This risk is highest for practical skills, soft skills, and project performance. A sensitivity model that removes these fields can determine whether a coherent association remains beyond the most criterion-proximal measures. A reduction in explained variance would indicate substantial shared information, but it would not partition genuine capability from common measurement. That distinction would require item-level construct mapping, independent or blinded raters, temporal separation, and a multitrait-multimethod design.

Converting a continuous 0-100 score into low, moderate, and high categories introduces further measurement loss. Categorization removes within-band variation, creates instability near thresholds, and may imply unsupported qualitative distinctions [46]. The continuous score should therefore be retained for association analysis, while classification errors should be evaluated by ordinal distance because adjacent and non-adjacent errors have different substantive consequences.

2.5. Machine-Learning Validation and Model Evaluation

Machine-learning models can represent nonlinearities and interactions, but they cannot repair an ambiguous target. High accuracy may arise because a classifier reproduces the competence structure, thresholds, and institutional practices used to generate administrative bands. Such performance demonstrates internal reproducibility within the archive; it does not establish prospective employment prediction, transportability, or external validity.

Internal estimates are vulnerable to leakage and model-selection bias. Leakage arises when validation or test information influences imputation, scaling, feature selection, tuning, or construction. At the same time, repeated optimization on the same validation data can produce optimistic performance even without explicit target leakage [18], [47]. Preprocessing must be estimated within training data, hyperparameters selected without test outcomes, transparent baselines retained, and final metrics calculated once in an untouched held-out partition [20], [21]. When observations are clustered, resampling must reflect the dependence structure and intended generalization claim [19].

No single metric adequately evaluates multiclass ordinal prediction. Accuracy can obscure minority-class failure, whereas macro-F1 weights class-specific precision and recall equally. One-versus-rest ROC-AUC and precision-recall AUC evaluate probability ranking, with precision-recall curves being particularly informative under imbalance [48]. Calibration should be examined through the Brier score, calibration intercept, calibration slope, and a defined calibration-error measure [25], [49]–[51]. Weighted kappa and absolute ordinal distance distinguish adjacent from non-adjacent errors [52]. Uncertainty intervals and paired comparisons are required before numerical differences are interpreted as algorithmic superiority.

Comparisons among ordinal or regularized regression, support-vector machines, random forests, gradient boosting, and neural networks should therefore be framed as internal comparative evaluation. Greater complexity does not guarantee better out-of-sample performance [51]. Transparent baselines may be preferable when discrimination is similar, but interpretability or calibration is stronger. TRIPOD+AI and PROBAST+AI support transparent reporting of development, validation, bias, and applicability [22], [24], but they do not convert an administrative rating into an externally validated endpoint.

2.6. Explainability, Subgroup Analysis and Governance

Permutation importance and SHAP can support model auditing by quantifying performance dependence and decomposing individual predictions [53], [54]. Agreement among standardized regression coefficients, permutation importance, and aggregated SHAP values may indicate a stable predictor ordering. These methods remain descriptive; they do not establish causal effects, construct validity, or intervention targets. Post-hoc explanations may instead create unwarranted confidence when target definition, provenance, or deployment context is weak [26], [28], [55].

Subgroup differences in accuracy, macro-F1, calibration, or ordinal error may reflect sample size, outcome prevalence, feature distributions, assessment practices, resources, or model instability. They do not constitute a fairness audit. Fairness assessment requires protected attributes, a prespecified fairness construct, adequate subgroup information, and explicit analysis of benefits, burdens, and decision consequences [29], [31],

[32], [56]. Because the four archived groups simultaneously represent institutions and tracks, descriptive variation cannot separate model behavior from contextual or measurement differences.

Educational prediction is a sociotechnical process. Bias may enter through construct definitions, historical labels, opportunity structures, feature availability, missingness, objectives, thresholds, interfaces, and human responses [56], [57]. Governance should therefore define purpose, acceptable and prohibited uses, human review, documentation, access controls, uncertainty communication, and mechanisms for correcting records. Model cards can formalize these boundaries [58]. Under the available evidence, classification is defensible only for low-stakes internal auditing, mentoring, curriculum review, and identification of records requiring additional review; it is not defensible for selection, certification, graduation, or restriction of opportunity.

3. MATERIALS AND METHODS

3.1. Study Design, Analytical Objectives and Reporting Framework

A cross-sectional secondary analysis was conducted using de-identified administrative records generated through routine end-of-programs monitoring during the 2023–2024 academic year. The readiness outcome and all learner-level indicators were recorded within the same administrative measurement window; no temporal separation existed between predictor measurement and outcome ascertainment. The analysis therefore estimated concurrent associations and the internal reproducibility of archived readiness categories. It did not constitute a prospective prediction study, an intervention evaluation, a quasi-experiment, a causal mediation analysis, or validation against realized labor-market outcomes.

Four prespecified analytical objectives structured the analysis. First, the cohort composition, missing-data pattern, continuous readiness distribution, and three administrative readiness bands were described. Second, pooled bivariate and multivariable associations were estimated, followed by a sensitivity analysis that excluded indicators with the greatest potential for criterion overlap. Third, six candidate classifiers were compared with a majority-class baseline in a single untouched internal test partition. Fourth, model reliance, robustness, and institution–track-specific outputs were examined as secondary diagnostics.

STROBE guided reporting for observational studies, TRIPOD+AI for prediction-model reporting, and PROBAST+AI for risk-of-bias and applicability assessment [22], [24]. PROBAST was also consulted to maintain continuity with established prediction-model appraisal principles [59]. These frameworks were adapted because they were developed primarily for epidemiological and clinical prediction research rather than vocational-education analytics. Their use structured the reporting of leakage control, validation, and interpretation; it did not convert the administrative readiness rating into a clinical or externally validated endpoint.

3.2. Setting, Data Source, Unit and Eligibility

The source archive included final-year learners from four anonymized polytechnic-style vocational institutions. Each institution contributed one vocational track—Engineering, Information Technology, Business Services, or Health-Support creating a one-to-one correspondence between institution and track. Consequently, between-track variation could not be separated from between-institution variation.

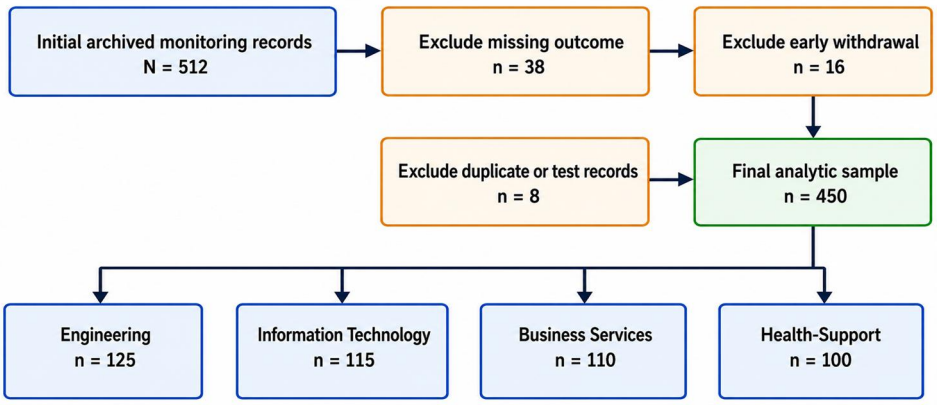

Figure 1. Record derivation and distribution of the final analytic sample across the four institution–track groups

The sampling frame comprised a census of 512 end-of-program monitoring records. Eligibility required a record to represent a final-year learner, contain a valid institution–track code, and include the archived readiness outcome at the end of the monitoring period. Exclusions were applied before imputation and modelling: 38 records lacked the archived readiness outcome, 16 represented withdrawals before end-of-cycle archiving, and eight were duplicate or test records. The final analytic sample consisted of all 450 eligible learner records. Figure 1

summarizes record derivation and the distribution of the analytic sample across the four institution–track groups.

No a priori sample-size calculation was performed because the complete eligible archive, rather than a sampled subset, was analyzed. Statistical precision was represented using confidence intervals and internal resampling. Census inclusion does not establish adequate information for stable subgroup, fairness, or external-validation claims; those claims depend on event counts, class balance, model complexity, and the intended use of predictions [60].

3.3. Outcome, Candidate Indicators and Measurement Boundaries

The primary outcome was an archived 0–100 employability-readiness composite generated by programmed staff from placement-supervisor judgement, capstone-defense information, and a final readiness rubric. The continuous rating was used in descriptive, correlational, and linear-regression analyses. Prespecified institutional thresholds were retained without re-estimation for classification: low (<60), moderate (60–79), and high (≥80). These categories represented administrative programmed-completion bands. They did not measure job acquisition, retention, earnings, occupational alignment, employer-rated productivity, or career progression.

Six candidate indicators were available. Practical skills, digital competence, soft skills, internship intensity, and project performance varied at the learner level. The industry–education integration index was recorded once for each institution and then assigned to every learner record within that institution. Although the index appeared in 450 rows, it contributed only four unique contextual values. Associations involving this index were therefore interpreted as record-weighted descriptive summaries rather than independent learner-level or institution-level effects.

Potential measurement overlap was specified before interpretation. Practical-skills, soft-skills, and project-performance ratings could share rubric content, assessors, assessment occasions, or settings with the readiness composite. Separate database fields were not treated as evidence of construct independence. Validity concerns the defensibility of an intended score interpretation, and common measurement sources may inflate observed associations [12], [13], [17]. Table 1 therefore distinguishes the recorded content of each variable from its permissible analytical interpretation.

Table 1. Operational definitions, scoring, and analytical boundaries of the archived variables (N = 450 units)

Variable	Role and scale	Archived content	Analytical role and boundary
Employability readiness	Outcome; 0–100; low <60, moderate 60–79, high ≥80	Placement-supervisor judgement, capstone defense, and final readiness rubric	Continuous outcome and derived classification target; not an independent employment outcome
Practical skills	Learner-level; 0–100	Task accuracy, safety, troubleshooting, and equipment handling	Association and classification indicator; possible rubric overlap
Digital competence	Learner-level; 0–100	Information processing, communication, content creation, safety, and problem solving	Association and classification indicator
Soft skills	Learner-level; 0–100	Communication, teamwork, professionalism, and adaptability	Association and classification indicator; possible rubric overlap
Internship intensity	Learner-level; 0–12 months	Duration of structured workplace learning	Association and classification indicator; duration does not measure placement quality
Project performance	Learner-level; 0–100	Quality, timeliness, collaboration, and applied implementation	Association and classification indicator; possible rubric overlap
Integration index	Institution-linked proxy; 0–100	Employer involvement, feedback, joint assessment, resources, and curriculum linkage	Exploratory indicator only; four values with complete institution–track confounding

3.4. Data Extraction and Integrity Checks

Eligibility was resolved before data transformation. Withdrawal, duplicate, and test records were identified from the archived status field. Range checks applied the documented limits of 0–100 to the readiness, competence, project, and integration scores and 0–12 months to internship intensity. Track codes were checked against the four admissible categories. No eligible record was removed solely because it contained an extreme but in-range predictor value. Potentially influential observations in the linear model were evaluated using Cook's distance rather than deleted automatically [23], [61].

The readiness outcome was complete after eligibility exclusions. Predictor missingness ranged from 1.3% to 3.1%. Little's MCAR test was used only to assess whether the observed pattern was inconsistent with missing completely at random; a non-significant result was not treated as proof of the MCAR mechanism [62]. For association analyses, the archived workflow used multiple

imputation by chained equations with 20 imputations and predictive mean matching for continuous variables, consistent with established guidance for flexible multivariable imputation [63], [64]. Pooled imputed estimates were compared with complete-case estimates as a robustness analysis.

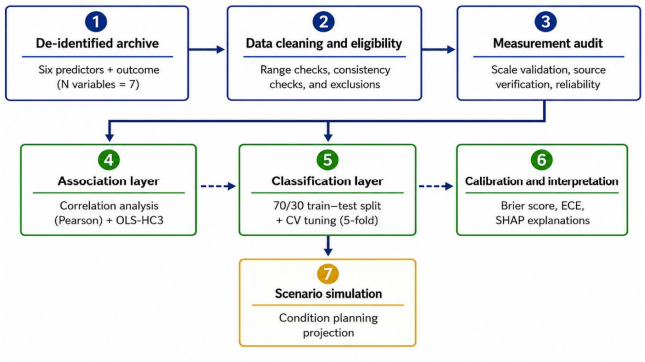

Figure 2. Analytical workflow and leakage-control logic. Preprocessing for supervised models is estimated only from training data or training folds.

Classification preprocessing was isolated from evaluation data. Imputation parameters were estimated within the development partition or the current cross-validation training fold and were then applied unchanged to the corresponding validation or test records. Standardization parameters were handled identically for algorithms requiring scaled inputs. Held-out outcomes were not used to estimate imputation, scaling, feature-processing, or tuning parameters, and readiness bands were not included as imputation predictors. This separation was imposed to prevent test-to-training information leakage, which can produce optimistic performance estimates [18].

3.5. Descriptive and Bivariate Association Analyses

Continuous variables were summarized using the mean, standard deviation, observed minimum, and observed maximum. Missingness was reported as a count and percentage for each variable. Institution–track composition and readiness bands were summarized using frequencies and percentages. The continuous readiness distribution was displayed in five-point intervals, with the prespecified administrative thresholds superimposed rather than estimated from the observed data.

Pearson product–moment correlations were used as the primary bivariate association estimates. Bias-corrected and accelerated 95% bootstrap confidence intervals were generated from 5,000 archived resamples to reduce dependence on normal-theory interval assumptions [65]. Family-wise multiplicity across the 21 unique pairwise correlations in the seven-variable matrix was controlled using Holm's sequential adjustment [66]. Spearman rank correlations were calculated as a sensitivity analysis for monotonic associations and potential non-normality. Correlations involving the integration index were retained only as descriptive record-weighted values because the index contained four independent contextual values.

3.6. Multivariable Association Model and Criterion-Overlap Sensitivity

Ordinary least squares regression was used to estimate the association between the continuous readiness rating and practical skills, digital competence, soft skills, internship intensity, project performance, and the integration index. The full model was specified as follows:

$$ER_i = \beta^0 + \beta^1 PS_i + \beta^2 DC_i + \beta^3 SS_i + \beta^4 II_i + \beta^5 PP_i + \beta^6 IEI_j + \varepsilon_i \quad (1)$$

In this expression, ER denotes the administrative readiness rating; PS, practical skills; DC, digital competence; SS, soft skills; II, internship intensity; PP, project performance; and IEI, the integration value assigned to institution j. Because IEI varied across only four institutions and institution was identical to track, β_6 was interpreted as a descriptive record-level coefficient rather than an independently identified contextual effect.

Inference used HC3 heteroskedasticity-consistent standard errors because of their improved finite-sample performance under leverage and heteroskedasticity [67]. Models reported R^2, adjusted R^2, omnibus F statistics, standardized coefficients, and variance-inflation factors. Residual normality, heteroskedasticity, and influential observations were assessed using the Shapiro–Wilk test, Breusch–Pagan test, and maximum Cook's distance, respectively, as descriptive diagnostics rather than automatic acceptance criteria.

A prespecified sensitivity model excluded practical skills, soft skills, and project performance because of their potential overlap with the outcome, while retaining digital competence, internship intensity, and the integration index. Changes in adjusted R^2 and coefficient ranking were examined. Any reduction in adjusted R^2 indicated sensitivity to removing overlapping information, not variance caused by contamination. Because the archived unstandardized coefficients were inconsistent with the reported means and the original output was unavailable, unstandardized slopes, intercepts, standard errors, and confidence intervals were omitted.

3.7. Data partitioning, leakage control and model development

The three classification bands contained 78 low-, 201 moderate-, and 171 high-readiness records. A stratified random split with seed 42 allocated 70% of records to development (n = 315: 55 low, 140 moderate, and 120 high) and 30% to an internal held-out test partition (n = 135: 23 low, 61 moderate, and 51 high). The test partition remained inaccessible during imputation fitting, standardization, hyperparameter selection, early

stopping, threshold modification, and model selection. It was used only for final evaluation.

Hyperparameters were tuned within the development partition using five-fold stratified cross-validation. Macro-F1 was the principal tuning criterion because it assigns equal influence to each class and reduces dominance by the moderate-readiness class. Preprocessing was re-estimated within every training fold. This fold-specific procedure prevented validation-fold information from influencing imputation or scaling [18]. Institution–track structure was not used to define cross-validation folds in the archived analysis. The held-out assessment therefore represents a random internal split rather than leave-one-institution-out or transportability validation. Structured resampling would be required to support claims about performance in a new institution [19].

3.8. Candidate classifiers and tuning specifications

The comparison panel comprised a majority-class reference and six candidate classifiers representing ordinal, regularized linear, kernel, ensemble-tree, boosting, and neural-network approaches. The proportional-odds model preserved the ordering of the outcome [68]. Elastic-net multinomial regression provided a regularized linear comparator. The radial-basis-function support-vector machine represented nonlinear decision boundaries in a scaled feature space [69]. Random Forest and XGBoost represented bagged and boosted tree ensembles, respectively [26]. The back-propagation neural network used two fully connected hidden layers with rectified linear activations, dropout, L2 regularization, and a SoftMax output layer, consistent with established neural-network regularization procedures [70], [71]. Table 2 specifies the retained architecture and hyperparameter grids.

Table 2. Candidate classification models and retained implementation specifications

Model	Preprocessing or structure	Archived tuning or fixed specification
Majority baseline	Most frequent development-set class	No tuning
Ordinal logistic	Standardized inputs; proportional-odds model	Maximum-likelihood estimation; proportional-odds assumption assessed
Elastic-net multinomial	Standardized inputs; multinomial loss	l1_ratio: 0.10, 0.50, 0.90; C: 0.01, 0.10, 1, 10; maximum iterations: 5,000
SVM (RBF)	Standardized inputs; balanced class weights	C: 0.10, 1, 10, 100; gamma: scale, 0.01, 0.10, 1
Random Forest	Fold-imputed inputs; scaling not required	Trees: 300–1,000; depth: none, 5, 10; minimum leaf: 1, 3, 5; maximum features: sqrt or log2
XGBoost	multi:softprob objective; fold-imputed inputs	Trees: 100–500; depth: 2–4; eta: 0.01–0.10; subsample and colsample: 0.70–1.00
BPNN	Six inputs; hidden layers 64 → 32; ReLU; dropout 0.20; L2 = 10^{-4}; SoftMax output	Adam optimizer; learning rate 10^{-3}; batch size 32; maximum 200 epochs; early-stopping patience 15; class-weighted loss

The revision archive did not retain the exact software packages and versions, cross-validation fold identifiers, class-weight values, initialization states beyond the global split seed, or complete tuning logs. The model families, tuning grids, split rule, and evaluation sequence can therefore be reconstructed analytically. Exact numerical reproduction requires the original software environment, row-level data, and saved resampling indices.

3.9. Held-out Validation, Ordinal Error and Calibration

Each candidate classifier was evaluated once in the 135-record held-out partition. Class-label performance was quantified using accuracy, macro precision, macro recall, macro-F1, and weighted-F1. Ranking performance was assessed using one-versus-rest macro-ROC-AUC and PR-AUC. PR-AUC was retained because precision–recall summaries can be informative when class frequencies are unequal [48]. Ordinal agreement was assessed using linearly weighted Cohen's κ and mean absolute ordinal distance (MAOD):

$$MAOD = (1/n)\,\Sigma\,|y_i - \hat{y}_i| \quad (2)$$

The three readiness bands were coded in their natural order as 0, 1, and 2. The confusion matrix was reported as counts and row percentages so that direct low-to-high errors could be distinguished from adjacent-band errors. Performance interpretation considered discrimination, calibration, and error severity rather than overall accuracy alone [23], [25], [51].

Archived 95% confidence intervals were based on 1,000 bootstrap resamples of the held-out set where retained values were available. Prediction-level files and the paired multiclass covariance implementation were unavailable. Pairwise significance tests between classifiers were therefore excluded, and point-estimate rankings were interpreted descriptively rather than as evidence of statistical superiority or equivalence.

Archived calibration outputs comprised a multiclass Brier score, calibration slope, calibration intercept, and expected calibration error. These quantities were interpreted only within the held-out partition. The archive did not retain probability-level predictions, class-specific

calibration components, Brier-score normalization, ECE binning rules, or calibration code. Recalculation, calibration plots, temporal or external calibration, and decision-curve analysis were therefore not claimed. This interpretive boundary is consistent with the requirement to distinguish discrimination from calibration and applicability [22], [25].

3.10. Candidate Model Reliance and Monitoring

For the classifier with the highest held-out point estimates, permutation importance was calculated as the reduction in held-out macro-F1 after permuting each predictor. At the same time, mean absolute SHAP values were used as a complementary summary of predictor contribution [54]. Both statistics were interpreted only as indicators of fitted-model dependence, not as causal effects, intervention targets, or evidence that changing a predictor would alter readiness. Post-hoc explanation cannot remedy an invalid or contaminated target, and interpretable baseline models remain necessary when predictions may inform consequential decisions [28].

Macro-F1 and class-specific true-positive rates were also summarized across the four institution–track groups. A maximum macro-F1 spread of 0.05 was treated as a descriptive signal for further investigation rather than a prespecified fairness threshold. Because subgroup cells were small, institution and track were inseparable, and protected attributes were unavailable, this analysis did not constitute a fairness audit or demonstrate the absence of bias [29], [31]. Responsible educational analytics additionally requires transparency, contestability, and evaluation of the consequences produced by labels and interventions [33].

4. RESULTS

4.1. Sample Derivation and Data Completeness

The analytic sample was derived from 512 end-of-program records. Before imputation and model estimation, 62 records were excluded: 38 did not contain an archived readiness outcome, 16 corresponded to learners who withdrew before end-of-cycle outcome archiving, and eight were duplicate or test entries. The resulting analytic file contained 450 eligible learner records. Records were distributed across Engineering (n = 125), Information Technology (n = 115), Business Services (n = 110), and Health-Support (n = 100). Because one institution contributed one track, institution and track were structurally confounded; the group labels therefore denote institution-track strata rather than independently estimable institutional or disciplinary effects.

Table 3. Data completeness and descriptive statistics for the analytic sample.

Variable	Missing (n)	Missing (%)	Mean	SD	Observed minimum	Observed maximum
Practical skills	6	1.30	76.35	14.28	35.00	98.00
Digital competence	9	2.00	68.42	16.53	25.00	95.00
Soft skills	7	1.60	71.18	11.95	40.00	94.00
Internship intensity (months)	14	3.10	6.40	2.10	0.00	12.00
Project performance	8	1.80	70.20	13.10	30.00	96.00
Integration-index linkage	10	2.20	65.80	15.42	20.00	90.00
Employability readiness	0	0.00	72.50	12.35	35.00	95.00

The readiness outcome had no missing values in the analytic file. Predictor missingness remained limited, with missing counts ranging from six records (1.30%) for practical skills to 14 records (3.10%) for internship intensity. Little's MCAR test did not reject the null pattern, $\chi^2(18) = 21.7$, $p = 0.24$ [62]. Consistent with the inferential scope of the test, this non-significant result was interpreted as an absence of statistical evidence against MCAR rather than confirmation of the missingness mechanism. Tables 3 and 4 report the variable-level completeness and descriptive summaries used in the subsequent analyses.

4.2. Cohort Profile and Readiness-band Distribution

The continuous archived readiness rating averaged 72.50 (SD = 12.35) and ranged from 35 to 95. Band classification placed 78 records (17.3%) in low readiness, 201 records (44.7%) in moderate readiness, and 171 records (38.0%) in high readiness. The combined moderate and high bands represented 372 records (82.7%), and the moderate band was the modal category. Figure 3 shows that the distribution crossed both predefined thresholds and was not concentrated within a single readiness segment.

Among learner-level indicators scored on the 0-100 scale, practical skills produced the highest mean (M = 76.35, SD = 14.28). Digital competence produced the lowest mean and the largest dispersion (M = 68.42, SD = 16.53). Soft skills averaged 71.18 (SD = 11.95), and project performance averaged 70.20 (SD = 13.10). Internship intensity averaged 6.40 months (SD = 2.10; observed range = 0-12). The record-weighted integration index averaged 65.80 (SD = 15.42), although this statistic reflected four institution-level values repeated across learner records. Institution-track representation was comparatively even, ranging from 22.2% to 27.8% of the cohort (Table 4 and 5).

Table 4. Distribution by institution–track group.

Institution–track group	N = 450	Percent (%)
Engineering	125	27.80
Information Technology	115	25.60
Business Services	110	24.40
Health-Support	100	22.20

Table 5. Distribution by administrative readiness band.

Readiness band	N = 450	Percent (%)
Low (<60)	78	17.30
Moderate (60–79)	201	44.70
High (≥80)	171	38.00

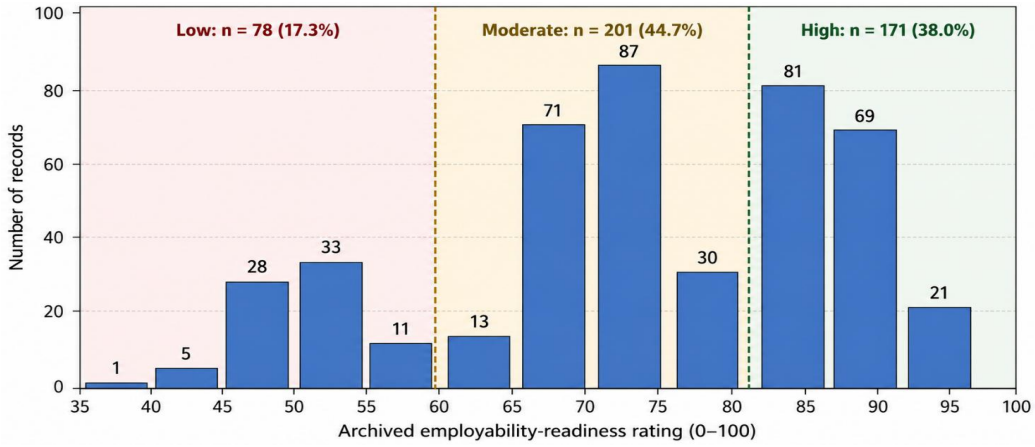

Figure 3. Distribution of the archived readiness rating. Bars reproduce the archived five-point frequency bins; background bands and dashed lines mark the predefined low (<60), moderate (60-79), and high (≥80) thresholds.

4.3. Bivariate Association Structure

All five learner-level indicators showed positive bivariate associations with the continuous administrative readiness rating, and every 95% BCa confidence interval excluded zero. Practical skills had the largest coefficient, r = 0.72, 95% BCa CI (0.67, 0.76), followed by digital competence, r = 0.65, 95% BCa CI (0.59, 0.70). Internship intensity and project performance showed similar coefficients, r = 0.61, 95% BCa CI (0.55, 0.66), and r = 0.60, 95% BCa CI (0.54, 0.66), respectively. Soft skills produced the smallest learner-level coefficient, r = 0.54, 95% BCa CI (0.47, 0.60).

Inter-indicator Pearson correlations ranged from 0.36 to 0.52. The largest coefficient was observed between practical skills and project performance (r = 0.52), while the remaining coefficients were below 0.50. This pattern indicates related but non-identical recorded indicators and aligns with the modest variance-inflation factors observed in the multivariable model. Spearman sensitivity estimates preserved the ordering of readiness associations and differed from the Pearson coefficients by no more than $|\Delta r| = 0.03$.

The record-weighted association between the integration index and readiness was r = 0.58, 95% BCa CI (0.52, 0.64). The apparent precision of this coefficient should not be interpreted as information from 450 independent contextual observations because the index contributed only four institution-level values and institution was completely confounded with track. Table 5 and Figure 4 therefore present the integration-index coefficients as descriptive archive features rather than institution-level effect estimates.

Table 6. Pearson correlation matrix with 95% BCa CI bootstrap confidence intervals.

Variable	PS	DC	SS	II	PP	IEI	ER
Practical Skills (PS)	1.00						
Digital Competence (DC)	0.48 [0.41, 0.55]	1.00					
Soft Skills (SS)	0.41 [0.33, 0.48]	0.39 [0.31, 0.47]	1.00				
Internship Intensity (II)	0.45 [0.37, 0.52]	0.42 [0.34, 0.49]	0.36 [0.28, 0.44]	1.00			
Project Performance (PP)	0.52 [0.45, 0.58]	0.46 [0.38, 0.53]	0.44 [0.36, 0.51]	0.40 [0.32, 0.47]	1.00		
Industry-education Integration Index (IEI)	0.43 [0.35, 0.50]	0.47 [0.39, 0.54]	0.38 [0.30, 0.46]	0.49 [0.42, 0.56]	0.41 [0.33, 0.48]	1.00	

Variable	PS	DC	SS	II	PP	IEI	ER
Employability Readiness (ER)	0.72 [0.67, 0.76]	0.65 [0.59, 0.70]	0.54 [0.47, 0.60]	0.61 [0.55, 0.66]	0.60 [0.54, 0.66]	0.58 [0.52, 0.64]	1.00

Note. Intervals were estimated from 5,000 archived resamples using the BCa method, which adjusts percentile limits for estimated bias and acceleration in the bootstrap distribution [72].

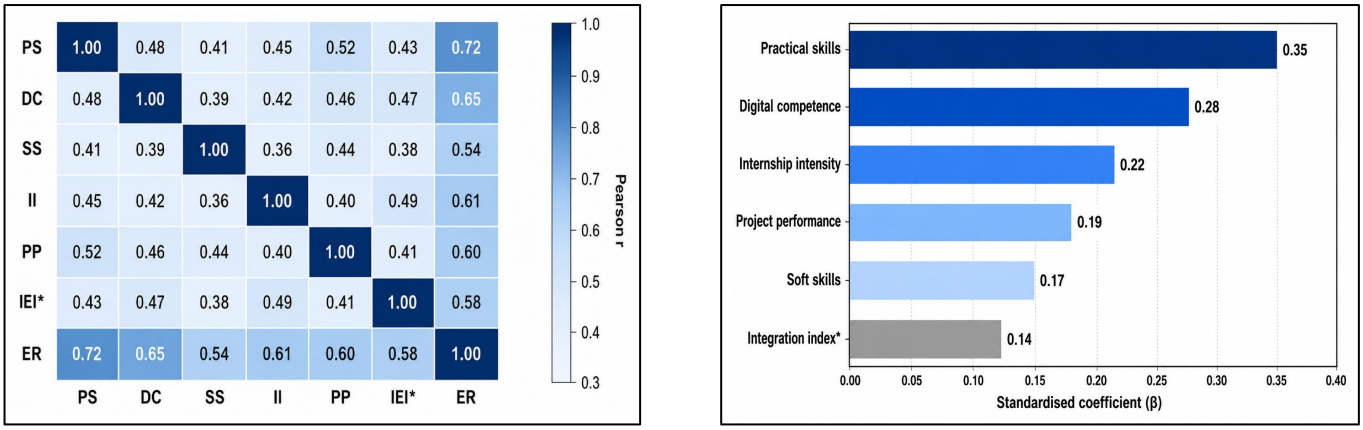

Figure 4. Bivariate and multivariable association structure. (A) displays the archived Pearson matrix; the dashed boundary marks the institution-linked integration index (IEI*); (B) displays standardized coefficients from the archived six-input model.

4.4. Multivariable Regression, Model Diagnostics and Criterion-Overlap Sensitivity Analysis

The full six-input ordinary least-squares model explained a substantial proportion of the observed variation in the concurrent readiness rating, $R^2 = 0.76$, adjusted $R^2 = 0.75$, $F(6, 443) = 234.8$, $p < 0.001$. Practical skills had the largest standardized coefficient ($\beta = 0.35$), followed by digital competence ($\beta = 0.28$), internship intensity ($\beta = 0.22$), project performance ($\beta = 0.19$), soft skills ($\beta = 0.17$), and the integration index ($\beta = 0.14$). Variance-inflation factors ranged from 1.49 to 1.86. This range indicates limited variance inflation in the fitted coefficient estimates, although VIFs were interpreted contextually rather than against a rigid universal cut-off [73].

Archived diagnostics reported Shapiro-Wilk ($W = 0.992$), ($p = 0.21$); Breusch-Pagan ($\chi^2 = 70.34$), ($p = 0.29$); and maximum Cook's ($D = 0.041$). Inference was based on HC3 covariance estimation, a leverage-adjusted heteroskedasticity-consistent procedure developed to improve finite-sample performance [67]. Taken together, the diagnostics did not identify an assumption violation or influential record sufficient to overturn the adjusted pattern. These diagnostics do not resolve the common measurement-window and potential criterion-overlap concerns described in the Methods section.

A prespecified criterion-overlap sensitivity model excluded practical skills, soft skills, and project performance because those indicators were most susceptible to shared rubric content or shared assessment processes with the readiness outcome. The reduced model retained adjusted $R^2 = 0.58$, representing an absolute decrease of 0.17 from the full model. Digital competence remained the leading retained coefficient ($\beta = 0.34$), followed by internship intensity ($\beta = 0.27$) and the integration index ($\beta = 0.21$). The retained ordering was coherent after the three overlap-prone indicators were removed, although the smaller explained proportion shows that the excluded fields contained substantial information about the archived rating. The decrease cannot be attributed entirely to contamination because the removed measures may also contain genuine capability information.

Table 7 links the full and reduced coefficient patterns to their permissible interpretations. Unstandardized slopes, intercepts, standard errors, and confidence intervals are not reported because the archived unstandardized table failed internal arithmetic reconciliation and the original model output was unavailable. The integration-index coefficient remains a record-level descriptive quantity rather than an independently identified contextual effect.

Table 7. Archived standardized coefficients and criterion-overlap sensitivity.

Predictor	B (unstd.)	S.E.	β (std.)	95% CI for B	p-value	VIF	Reduced-model β
Practical Skills	0.30	0.030	0.35	[0.24, 0.36]	< 0.001	1.86	—
Digital Competence	0.21	0.027	0.28	[0.16, 0.26]	< 0.001	1.74	0.34
Internship Intensity	1.29	0.22	0.22	[0.86, 1.72]	< 0.001	1.55	0.27
Project Performance	0.18	0.031	0.19	[0.12, 0.24]	< 0.001	1.62	—

Predictor	B (unstd.)	S.E.	β (std.)	95% CI for B	p-value	VIF	Reduced-model β
Soft Skills	0.18	0.034	0.17	[0.11, 0.25]	< 0.001	1.49	—
Integration Index	0.11	0.029	0.14	[0.05, 0.17]	< 0.001	1.71	0.21
Model Summary	**R^2 = 0.76**	**Adj. R^2 = 0.75**	**F(6, 443) = 234.8**	**p < 0.001**	**Shapiro–Wilk W = 0.991**	**BP test p = 0.18**	**Adj. R^2 = 0.58**

Practical skills showed the largest adjusted association in the full model but, together with project performance and soft skills, was excluded from the reduced model because of potential criterion overlap. Digital competence and internship intensity retained positive coefficients, indicating associations that remained after removal of overlap-prone indicators. The integration index also remained positive but should be interpreted as exploratory because of its limited variation and complete confounding with institution–track membership. Accordingly, the coefficients represent concurrent within-system associations rather than causal or externally generalizable effects.

4.5. Held-out Classification Performance and Ordinal Agreement

The stratified partition allocated 315 records to model development and retained 135 records for a single held-out evaluation. The held-out set contained 23 low-readiness, 61 moderate-readiness, and 51 high-readiness records. Every candidate classifier exceeded the majority-class baseline on accuracy and macro-F1. Candidate-model accuracy ranged from 0.848 for ordinal logistic regression to 0.904 for the BPNN, compared with 0.452 for the baseline; macro-F1 ranged from 0.83 to 0.903, compared with 0.207 for the baseline. The improvement over the baseline was observed across ordinal, regularized linear, kernel, tree-ensemble, boosting, and neural-network model families. Because held-out class frequencies were unequal, macro-F1 was interpreted alongside accuracy so that performance in the largest class did not dominate the summary; this complementary reporting is consistent with established principles for evaluating discrimination under class imbalance [23], [48].

The BPNN produced the highest point estimates: 122 of 135 records were classified correctly, yielding accuracy = 0.904, macro precision = 0.90, macro recall = 0.906, macro-F1 = 0.903, and weighted-F1 = 0.904. Its macro-ROC-AUC was 0.94, and its PR-AUC was 0.92. Weighted κ was 0.872, and mean absolute ordinal distance was 0.096 band units. Weighted κ incorporates the ordered severity of disagreement, allowing adjacent-band errors to contribute less than extreme-band errors under the specified weighting scheme [52]. XGBoost ranked second on the principal point estimates, with accuracy = 0.885 and macro-F1 = 0.87; its accuracy was 0.019 lower and its macro-F1 was 0.033 lower than the BPNN point estimates.

The archived 95% bootstrap interval for BPNN accuracy was (0.853, 0.948), compared with (0.832, 0.933) for XGBoost. These intervals overlapped substantially. Prediction-level outputs required for a paired comparison were not retained, and intervals were unavailable for several models. Model rankings based on a finite evaluation sample can be sensitive to selection variance and model-selection overfitting [47]. Accordingly, Table 8 and Figure 4 support a descriptive point-estimate ranking, not a statistical claim that the BPNN was superior, equivalent, or uniquely optimal.

Table 8. Held-out model performance and ordinal error.

Model	Accuracy	Macro-F1	Weighted-F1	Macro ROC-AUC	PR-AUC	Weighted κ	Ordinal distance
Majority baseline	0.452	0.207	0.281	0.50	NR	0.000*	0.548*
Ordinal logistic	0.848	0.83	0.85	0.90	0.86	0.79	0.18
Random Forest	0.854	0.84	0.85	0.89	0.86	0.82	0.16
Elastic net	0.859	0.85	0.86	0.91	0.87	0.81	0.18
SVM (RBF)	0.878	0.86	0.87	0.91	0.88	0.84	0.13
XGBoost	0.885	0.87	0.88	0.92	0.89	0.86	0.10
BPNN	0.904	0.903	0.904	0.94	0.92	0.872	0.096

Note. N = 135. NR = not retained because the PR-AUC averaging convention for the constant baseline was unresolved. (*) Baseline κ and ordinal distance were deterministically derived from the held-out class counts and the rule assigning every record to the moderate class. Archived 95% bootstrap intervals were BPNN accuracy (0.853, 0.948), BPNN macro-F1 (0.85, 0.94), BPNN macro-ROC-AUC (0.90, 0.97), XGBoost accuracy (0.832, 0.933), and XGBoost macro-ROC-AUC (0.88, 0.96).

4.6. BPNN Error Pattern and Archived Calibration

The BPNN correctly classified 21 of 23 low-readiness records (91.3%), 55 of 61 moderate-readiness records (90.2%), and 46 of 51 high-readiness records (90.2%). Class-specific F1 values were 0.894 for low readiness, 0.894 for moderate readiness, and 0.920 for high readiness. The proximity of these values indicates that the

aggregate macro-F1 was not driven solely by the largest class, namely the moderate-readiness band.

All 13 misclassifications occurred between adjacent readiness bands. Two low-readiness records were classified as moderate; three moderate-readiness records were classified as low and three as high; and five high-readiness records were classified as moderate. No low-readiness record was classified as high, and no high-readiness record was classified as low. This error pattern accounts for the small mean ordinal distance and indicates ordinal coherence within the held-out partition, without validating the institutional cut-points themselves (Table 9; Figure 5).

Table 9. BPNN held-out confusion matrix and class-specific performance.

True band	Predicted low	Predicted moderate	Predicted high	Precision	Recall	F1
Low (n = 23)	21 (91.3%)	2 (8.7%)	0 (0.0%)	0.875	0.913	0.894
Moderate (n = 61)	3 (4.9%)	55 (90.2%)	3 (4.9%)	0.887	0.902	0.894
High (n = 51)	0 (0.0%)	5 (9.8%)	46 (90.2%)	0.939	0.902	0.920

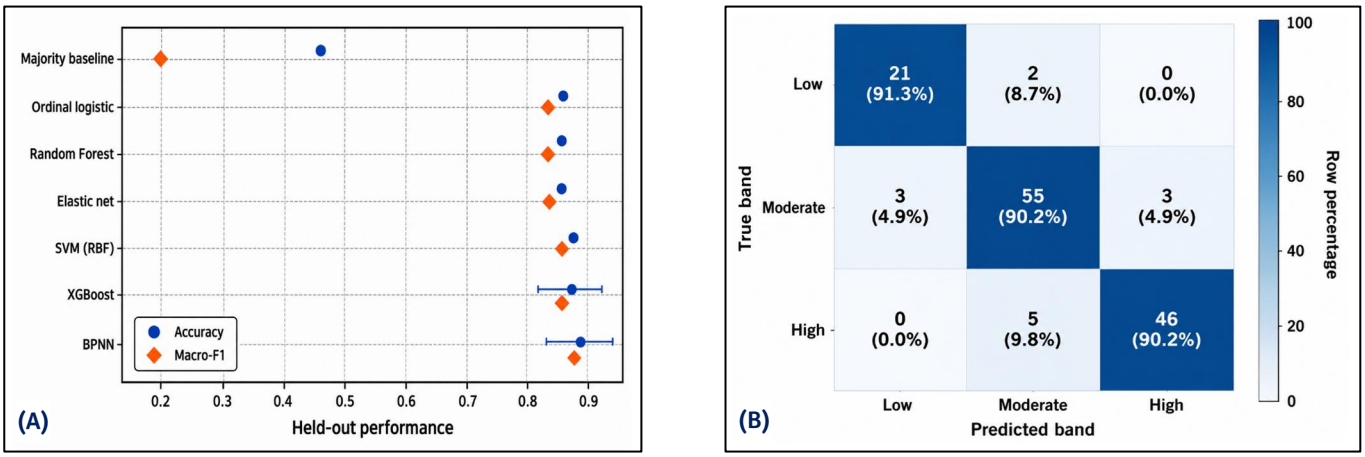

Figure 5. Held-out candidate-model performance and BPNN error pattern. (A) accuracy and macro-F1 point estimates; accuracy intervals are drawn only for XGBoost and BPNN because no other archived intervals were available; (B) BPNN counts and row percentages.

The archived calibration summary reported a multiclass Brier score of 0.092, a calibration slope of 0.98, a calibration intercept of 0.01, and an expected calibration error of 0.024. The Brier score summarizes overall probabilistic accuracy as a proper scoring rule [49], while the calibration intercept and slope jointly characterize weak calibration [50]. Numerically, the archived point estimates are compatible with close apparent calibration in the held-out set, consistent with the principle that discrimination and calibration should be reported as distinct performance domains [23]. They were not independently auditable because probability-level predictions, class-specific components, the multiclass Brier normalization, and the ECE binning convention were not retained. The calibration results are therefore reported as archived summaries only and do not establish calibration in another cohort, external setting, or decision context.

4.7. Model Reliance and Institution–Track Stability

Permutation importance and mean absolute SHAP values produced the same predictor ordering for the fitted BPNN. Permutation importance measures the reduction in predictive performance after a predictor is disrupted [74], whereas SHAP decomposes fitted predictions into additive feature-attribution values [49]. Practical skills ranked first, with a 0.064 reduction in held-out macro-F1 after permutation and mean |SHAP| = 0.182. Digital competence ranked second (0.052; 0.151), followed by internship intensity (0.041; 0.118), project performance (0.034; 0.098), soft skills (0.028; 0.082), and the integration index (0.021; 0.061). Agreement between the two summaries strengthens the description of which archived inputs the fitted model used. However, variable-importance estimates remain model-specific and do not identify causal determinants or justified intervention targets [53] integration-index result remains additionally constrained by its four repeated contextual values.

Held-out macro-F1 ranged from 0.87 to 0.92 across the four institution-track groups: Engineering, 0.91 (n = 38); Information Technology, 0.92 (n = 34); Business Services, 0.87 (n = 33); and Health-Support, 0.89 (n = 30). The maximum observed spread was 0.05. Across groups, true-positive rates ranged from 0.87 to 0.95 for low readiness, 0.86 to 0.92 for moderate readiness, and 0.89 to 0.93 for high readiness. Transparent model reporting supports disaggregated performance summaries while requiring explicit documentation of the evaluation setting,

intended use, and unsupported applications [58]. The observed subgroup differences therefore describe only this single random test partition; each subgroup was small, and each track represented a different institution. The display is a descriptive stability check, not evidence of fairness, absence of bias, transportability, or deployment readiness (Table 10; Figure 6).

Table 10. Archived BPNN model-reliance outputs.

Predictor	Permutation importance (macro-F1 reduction)	Mean |SHAP|	Rank
Practical skills	0.064	0.182	1st
Digital competence	0.052	0.151	2nd
Internship intensity	0.041	0.118	3rd
Project performance	0.034	0.098	4th
Soft skills	0.028	0.082	5th
Integration index	0.021	0.061	6th

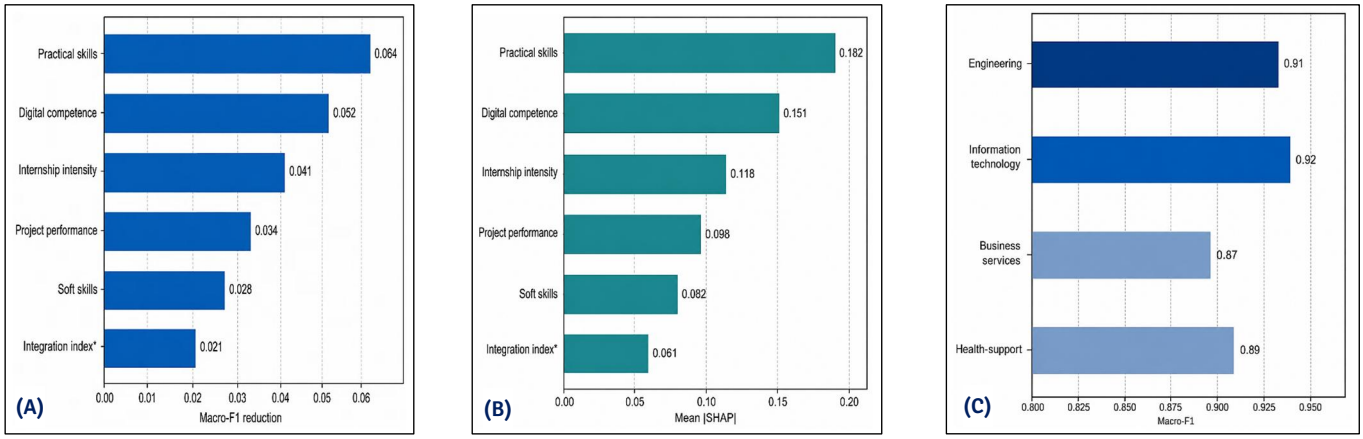

Figure 6. Archived BPNN interpretation and descriptive institution-track stability. (A) and (B) Permutation-related macro-F1 reduction and mean absolute SHAP values; (C) Held-out macro-F1 by institution-track group. The integration index (*) is context-linked, and the subgroup display is not a fairness audit.

5. DISCUSSION

The program-completion readiness rating was reproduced with substantial accuracy from indicators generated within the same assessment system. All learner-level indicators were positively associated with the score, with practical skills and digital competence producing the strongest correlations. The full model explained 76% of variance, whereas the reduced model retained an adjusted R^2 of .58 after practical skills, soft skills, and project performance were removed because of criterion-overlap concerns. All classifiers exceeded the baseline. In the held-out partition, the back-propagation neural network classified 122 of 135 records correctly, with an accuracy of 0.904 and a macro-F1 of 0.903. All errors occurred between adjacent categories, indicating preservation of ordinal structure.

The indicators form a coherent concurrent association structure, while digital competence and internship intensity retain explanatory information after exclusion of the predictors closest to the criterion. The findings do not establish employability in the broader labor-market sense, which includes capabilities, adaptability, career identity, human capital, institutional opportunity, and labor-market conditions [4], [5], [36], [75]. The analysis therefore concerns reproduction of archived administrative construct rather than employment attainment, retention, employer-rated performance, or causal effects.

The outcome combined placement-supervisor judgement, capstone-defense information, and an internal readiness rubric into three categories. Score validity requires evidence concerning content, response processes, internal structure, external relations, generalizability, and consequences of use [12], [13]. The analysis provides evidence of concurrent association and internal reproducibility, but not comprehensive employability validity. Practical skills, soft skills, and project performance may represent relevant capabilities while sharing raters, content, or assessment occasions with the outcome. The reduction in adjusted R^2 from 0.75 to 0.58 after their exclusion is consistent with shared information, although it cannot distinguish substantive convergence from common-method effects or criterion contamination [43]–[45].

Practical skills showed the strongest overall association, including the largest correlation and full-model standardized coefficient. This pattern accords with

vocational education's emphasis on demonstrated performance, but possible rubric overlap limits interpretation. Digital competence was more robust under sensitivity analysis. It retained the largest learner-level coefficient in the reduced model and remained highly ranked in permutation importance and mean absolute SHAP. This association is theoretically plausible because digital competence integrates information management, communication, collaboration, problem solving, and adaptability [40]. Internship intensity also remained positively associated with readiness, although duration alone does not capture placement quality, supervision, task authenticity, learner selection, or prior advantage [14], [15].

The integration index contained four institution-level values repeated across 450 records, while institution was perfectly confounded with vocational track. Its coefficients and importance values therefore describe four institution–track contexts rather than independent institutional effects. Categorization of the readiness score also removed within-band information and increased instability near the 60 and 80 cut-points [46]. Adjacent-category errors support ordinal coherence but do not validate these thresholds as educationally optimal or equitable.

Several non-exclusive explanations remain compatible with the observed pattern. The indicators may represent convergent vocational capabilities, but three predictors may also share information with the criterion. Converting a continuous score into three ordered categories may have simplified classification. Contextual structure may have entered through institution, track, and assessment culture. Exclusion of records without archived outcomes and early withdrawals may also have narrowed the analytic spectrum. These explanations require temporally ordered measurements, independent raters, complete cohort follow-up, and external outcomes.

The neural network produced the highest point estimates, but its 0.019 accuracy advantage over XGBoost does not establish statistical superiority. The development and test samples contained 315 and 135 records, prediction-level outputs were unavailable, and confidence intervals overlapped. Flexible models may produce optimistic estimates when tuning, selection, and evaluation are incompletely separated or when model complexity is high relative to sample size [18], [22], [59], [60]. Machine-learning methods do not consistently outperform regression under low risk of bias [51]. Algorithm selection should therefore prioritize validated incremental performance, stability, interpretability, and decision consequences [76]. When performance is comparable, ordinal or regularized models may be preferable because their parameters and failure modes are more auditable.

The neural network nevertheless demonstrated strong ordinal agreement. All 13 errors crossed one category boundary, weighted κ was 0.872, and mean absolute ordinal distance was 0.096. Several errors may therefore concern records near administrative thresholds rather than major ranking failures. Future studies should compare continuous, ordinal, and nominal formulations and report threshold distance, uncertainty, and borderline-decision consequences. Random record-level splitting is insufficient for transportability assessment because the same curriculum, assessors, and scoring culture may appear in both development and test data [19].

The statistics archived indicate low apparent calibration error in the held-out partition. These estimates remain provisional because probability-level predictions, class-specific components, binning procedures, and uncertainty intervals were unavailable. Calibration must be assessed separately from discrimination because stable ranking can coexist with inaccurate probabilities [23], [51].

Permutation importance and mean absolute SHAP ranked practical skills and digital competence as the main sources of model reliance. These measures do not identify causal determinants or intervention targets because they depend on the fitted model, predictor dependence, and background distribution [54], [55]. The model is suitable only for low-stakes auditing, threshold-sensitivity analysis, and secondary review. It should not independently determine progression, placement access, learner ranking, or employability status.

The institution–track macro-F1 range of 0.87–0.92 does not establish fairness. Each subgroup contained only 30–38 held-out records; institution and track were inseparable, protected and intersectional groups were not assessed, and estimates came from one partition. Similar aggregate performance can coexist with unequal error rates, calibration, opportunity access, or error consequences. Fairness assessment should define the intended decision, affected groups, plausible harms, and remedies, then report subgroup sensitivity, specificity, predictive values, calibration, and ordinal error with uncertainty [29], [57].

The study used the complete eligible archive, separated association from classification, restricted preprocessing to development data, retained a held-out partition, and compared multiple algorithms with a baseline. Sensitivity analyses showed stable coefficient directions and limited variation across imputation strategies. These procedures strengthen analytical assessment.

The principal limitations concern construct validity, causality, selection, transportability, fairness, and computational reproducibility. The outcome was not independently validated against labor-market criteria; several predictors may overlap with the criterion, institution and track were confounded, and excluded outcomes may have introduced selection bias. Missing software versions, fold assignments, prediction-level outputs, model logs, and fitted objects also restrict exact

reproduction. Future research should use temporally ordered measures, independent raters, explicit scoring rules, multiple institutions and tracks, and externally verified outcomes such as employment, retention, job–qualification match, workplace performance, and job quality. Model development should be preregistered and evaluated through nested resampling, optimism correction, temporal validation, and external validation. Until validity, transportability, calibration, decision utility, governance, and subgroup safety are demonstrated, the model should remain an audit instrument rather than a consequential decision tool.

6. CONCLUSION

The archived readiness rating was strongly associated with concurrently recorded competency indicators and could be reproduced within the same administrative system. The full model explained 76% of the variance, with an adjusted R^2 of 0.75, while the reduced model retained an adjusted R^2 of 0.58 after removing indicators most susceptible to criterion overlap. In the held-out partition, the BPNN achieved 0.904 accuracy and 0.903 macro-F1, with all 13 errors occurring between adjacent readiness categories. These results indicate a coherent internal scoring structure but do not establish employment prediction, causal effects, or neural-network superiority.

The model is therefore appropriate only for low-stakes auditing of scoring consistency, data quality, and borderline records. Consequential use would require prospective multi-institutional validation, independent employment and workplace outcomes, complete computational provenance, class-specific calibration, decision-utility assessment, and adequately powered fairness analyses. The principal contribution is the distinction between internal predictability and external validity: a readiness rubric may be highly reproducible within its originating system without being suitable for labor-market prediction or automated educational decisions.

ACKNOWLEDGMENTS

The authors would like to express their deepest gratitude to the Rajamangala University of Technology Krungthep, participants and colleagues for their exceptional support and resources, which were critical in facilitating this research.

CONFLICTS OF INTEREST

The authors declare that no conflicts of interest are associated with this study. All aspects of the research were conducted with the utmost integrity and transparency.

DATA AVAILABILITY

The datasets utilized and analyzed during this research are available from the corresponding author upon reasonable request.

ETHICAL STATEMENTS

The authors confirm that the study complied with all applicable local laws, ethical standards, and institutional guidelines, including obtaining approval from relevant ethics committees.

FUNDING

This research was conducted without financial support. The authors confirm that no funding was received for this study's research, analysis, or publication.

REFERENCES

[1] S. Jeon, How Can Innovative Technologies Transform Vocational Education and Training: Insights for Ukraine. Paris, France: OECD Publishing, 2025, https://doi.org/10.1787/fb40f416-en

[2] N. Kholifah et al., "Unlocking workforce readiness through digital employability skills in vocational education Graduates: A PLS-SEM analysis based on human capital Theory," Soc. Sci. Humanit. Open, vol. 11, p. 101625, 2025, https://doi.org/10.1016/j.ssaho.2025.101625

[3] M. Hardini, S. A. Anjani, S. Triandari, F. Amelia, and M. Rodriguez, "Orchestrating Big Data and Artificial Intelligence for Adaptive Digital Business Strategy," J. Comput. Sci. Technol. Appl., vol. 3, no. 2, pp. 185–198, 2026, https://doi.org/10.33050/qx8e0j55

[4] M. Clarke, "Rethinking graduate employability: the role of capital, individual attributes and context," Stud. High. Educ., vol. 43, no. 11, pp. 1923–1937, 2018, https://doi.org/10.1080/03075079.2017.1294152

[5] A. Eimer and C. Bohndick, "Employability models for higher education: A systematic literature review and analysis," Soc. Sci. Humanit. Open, vol. 8, no. 1, p. 100588, 2023, https://doi.org/10.1016/j.ssaho.2023.100588

[6] M. Tomlinson, "Forms of graduate capital and their relationship to graduate employability," Educ. Train., vol. 59, no. 4, pp. 338–352, 2017, https://doi.org/10.1108/ET-05-2016-0090

[7] H. Tushar and N. Sooraksa, "Global employability skills in the 21st century workplace: A semi-systematic literature review," Heliyon, vol. 9, no. 11, 2023, https://doi.org/10.1016/j.heliyon.2023.e21023

[8] K. Peersia, N. A. Rappa, and L. B. Perry, "Work readiness: definitions and conceptualisations," High. Educ. Res. Dev., vol. 43, no. 8, pp. 1830–1845, 2024, https://doi.org/10.1080/07294360.2024.2366322

[9] J. Borg, C. M. Scott-Young, and T. Bartram, "A Review of Graduate Work Readiness Literature: A Conceptual Exploration of the Implications for HRM Research and Practice," Asia Pacific J. Hum. Resour., vol. 63, no. 2, p. e70012, 2025, https://doi.org/10.1111/1744-7941.70012

[10] C. L. Caballero et al., "Untangling graduate employability and work readiness to inform evidence-based practice in higher education," High. Educ. Res. Dev., vol. 45, no. 5, pp. 1247–1264, 2026, https://doi.org/10.1080/07294360.2026.2615303

[11] D. Jackson, "Re-conceptualising graduate employability: The importance of pre-professional identity," High. Educ. Res. Dev., vol. 35, no. 5, pp. 925–939, 2016, https://doi.org/10.1080/07294360.2016.1139551

[12] S. Messick, "Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning," Am.

Psychol., vol. 50, no. 9, pp. 741–749, 1995, https://doi.org/10.1037/0003-066X.50.9.741
[13] P. M. Skorupiński, "American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Standards for educational and psychological testing," Kwart. Pedagog., vol. 60, no. 4, pp. 201–203, 2015, [Online]. Available: https://bibliotekanauki.pl/articles/892305
[14] D. Jackson and B. A. Dean, "The contribution of different types of work-integrated learning to graduate employability," High. Educ. Res. Dev., vol. 42, no. 1, pp. 93–110, 2023, https://doi.org/10.1080/07294360.2022.2048638
[15] P. M. L. Ng, T. M. Wut, and J. K. Y. Chan, "Enhancing perceived employability through work-integrated learning," Educ. Train., vol. 64, no. 4, pp. 559–576, 2022, https://doi.org/10.1108/ET-12-2021-0476
[16] R. Scandurra, D. Kelly, S. Fusaro, R. Cefalo, and K. Hermannsson, "Do employability programmes in higher education improve skills and labour market outcomes? A systematic review of academic literature," Stud. High. Educ., vol. 49, no. 8, pp. 1381–1396, 2024, https://doi.org/10.1080/03075079.2023.2265425
[17] P. M. Podsakoff, S. B. MacKenzie, J. Y. Lee, and N. P. Podsakoff, "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies," J. Appl. Psychol., vol. 88, no. 5, pp. 879–903, 2003, https://doi.org/10.1037/0021-9010.88.5.879
[18] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, 2023, https://doi.org/10.1016/j.patter.2023.100804
[19] D. R. Roberts et al., "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," Ecography (Cop.)., vol. 40, no. 8, pp. 913–929, 2017, https://doi.org/10.1111/ecog.02881
[20] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices," Int. J. Educ. Technol. High. Educ., vol. 17, no. 1, p. 3, 2020, https://doi.org/10.1186/s41239-020-0177-7
[21] M. Kuhn and K. Johnson, Applied predictive modeling, vol. 26. Springer, 2013, https://doi.org/10.1007/978-1-4614-6849-3
[22] X. Wang et al., "Chinese interpretation of PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," Chinese J. Clin. Thorac. Cardiovasc. Surg., vol. 32, no. 12, pp. 1686–1695, 2025, https://doi.org/10.7507/1007-4848.202507088
[23] E. W. Steyerberg et al., "Assessing the performance of prediction models: A framework for traditional and novel measures," Epidemiology, vol. 21, no. 1, pp. 128–138, 2010, https://doi.org/10.1097/EDE.0b013e3181c30fb2
[24] G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," Bmj, vol. 385, 2024, https://doi.org/10.1136/bmj-2023-078378
[25] B. Van Calster et al., "Calibration: The Achilles heel of predictive analytics," BMC Med., vol. 17, no. 1, p. 230, 2019, https://doi.org/10.1186/s12916-019-1466-7
[26] H. Khosravi et al., "Explainable Artificial Intelligence in education," Comput. Educ. Artif. Intell., vol. 3, p. 100074, 2022, https://doi.org/10.1016/j.caeai.2022.100074
[27] G. Ramaswami, T. Susnjak, A. Mathrani, and R. Umer, "Use of Predictive Analytics within Learning Analytics Dashboards: A Review of Case Studies," Technol. Knowl. Learn., vol. 28, no. 3, pp. 959–980, 2023, https://doi.org/10.1007/s10758-022-09613-x
[28] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, no. 5, pp. 206–215, 2019, https://doi.org/10.1038/s42256-019-0048-x
[29] R. S. Baker and A. Hawn, "Algorithmic Bias in Education," Int. J. Artif. Intell. Educ., vol. 32, no. 4, pp. 1052–1092, 2022, https://doi.org/10.1007/s40593-021-00285-9
[30] J. A. Idowu, "Debiasing Education Algorithms," Int. J. Artif. Intell. Educ., vol. 34, no. 4, pp. 1510–1540, 2024, https://doi.org/10.1007/s40593-023-00389-4
[31] M. Khalil, P. Prinsloo, and S. Slade, "Fairness, Trust, Transparency, Equity, and Responsibility in Learning Analytics," J. Learn. Anal., vol. 10, no. 1, pp. 1–7, 2023, https://doi.org/10.18608/jla.2023.7983
[32] R. F. Kizilcec and H. Lee, "Algorithmic fairness in education," in The Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates, Routledge, 2022, pp. 174–202. https://doi.org/10.4324/9780429329067-10
[33] S. Slade and P. Prinsloo, "Learning Analytics: Ethical Issues and Dilemmas," Am. Behav. Sci., vol. 57, no. 10, pp. 1510–1529, 2013, https://doi.org/10.1177/0002764213479366
[34] B. Williamson and R. Eynon, "Historical threads, missing links, and future directions in AI in education," Learning, Media and Technology, vol. 45, no. 3. Taylor & Francis, pp. 223–235, 2020. https://doi.org/10.1080/17439884.2020.1798995
[35] A. Mathrani, T. Susnjak, G. Ramaswami, and A. Barczak, "Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics," Comput. Educ. Open, vol. 2, p. 100060, 2021, https://doi.org/10.1016/j.caeo.2021.100060
[36] M. Fugate, A. J. Kinicki, and B. E. Ashforth, "Employability: A psycho-social construct, its dimensions, and applications," J. Vocat. Behav., vol. 65, no. 1, pp. 14–38, 2004, https://doi.org/10.1016/j.jvb.2003.10.005
[37] R. L. Brennan, "Commentary on 'Validating the Interpretations and Uses of Test Scores,'" J. Educ. Meas., vol. 50, no. 1, pp. 74–83, 2013, https://doi.org/10.1111/jedm.12001
[38] A. Rausch et al., "Designing an International Large-Scale Assessment of Professional Competencies and Employability Skills: Emerging Avenues and Challenges of OECD's PISA-VET," Vocat. Learn., vol. 17, no. 3, pp. 393–432, 2024, https://doi.org/10.1007/s12186-024-09347-0
[39] R. Vuorikari, S. Kluzer, and Y. Punie, DigComp 2.2: The Digital Competence Framework for Citizens—With New Examples of Knowledge, Skills and Attitudes. Luxembourg: Publications Office of the European Union, 2022, https://doi.org/10.2760/115376
[40] E. van Laar, A. J. A. M. van Deursen, J. A. G. M. van Dijk, and J. de Haan, "The relation between 21st-century skills and digital skills: A systematic literature review," Comput. Human Behav., vol. 72, pp. 577–588, 2017, https://doi.org/10.1016/j.chb.2017.03.010
[41] C. Succi and M. Canovi, "Soft skills to enhance graduate employability: comparing students and employers'

perceptions," Stud. High. Educ., vol. 45, no. 9, pp. 1834–1847, 2020, https://doi.org/10.1080/03075079.2019.1585420

[42] W. M. Adegbite, "Unpacking mediation and moderating effect of digital literacy and life-career knowledge in the relationship between work-integrated learning and graduate employability," Soc. Sci. Humanit. Open, vol. 10, p. 101161, 2024, https://doi.org/10.1016/j.ssaho.2024.101161

[43] D. T. Campbell and D. W. Fiske, "Convergent and discriminant validation by the multitrait-multimethod matrix," Psychol. Bull., vol. 56, no. 2, pp. 81–105, 1959, https://doi.org/10.1037/h0046016

[44] P. M. Podsakoff, S. B. MacKenzie, and N. P. Podsakoff, "Sources of method bias in social science research and recommendations on how to control it," Annu. Rev. Psychol., vol. 63, no. 1, pp. 539–569, 2012, https://doi.org/10.1146/annurev-psych-120710-100452

[45] J. F. Binning and G. V. Barrett, "Validity of Personnel Decisions: A Conceptual Analysis of the Inferential and Evidential Bases," J. Appl. Psychol., vol. 74, no. 3, pp. 478–494, 1989, https://doi.org/10.1037/0021-9010.74.3.478

[46] D. G. Altman and P. Royston, "The cost of dichotomising continuous variables," Br. Med. J., vol. 332, no. 7549, p. 1080, 2006, https://doi.org/10.1136/bmj.332.7549.1080

[47] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," J. Mach. Learn. Res., vol. 11, pp. 2079–2107, 2010, [Online]. Available: https://www.jmlr.org/papers/v11/cawley10a.html

[48] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS One, vol. 10, no. 3, p. e0118432, 2015, https://doi.org/10.1371/journal.pone.0118432

[49] G. W. Brier, "Verification of forecasts expressed in terms of probability," Mon. Weather Rev., vol. 78, no. 1, pp. 1–3, 1950, https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2

[50] B. Van Calster, D. Nieboer, Y. Vergouwe, B. De Cock, M. J. Pencina, and E. W. Steyerberg, "A calibration hierarchy for risk models was defined: From utopia to empirical data," J. Clin. Epidemiol., vol. 74, pp. 167–176, 2016, https://doi.org/10.1016/j.jclinepi.2015.12.005

[51] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," J. Clin. Epidemiol., vol. 110, pp. 12–22, 2019, https://doi.org/10.1016/j.jclinepi.2019.02.004

[52] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," Psychol. Bull., vol. 70, no. 4, pp. 213–220, 1968, https://doi.org/10.1037/h0026256

[53] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," J. Mach. Learn. Res., vol. 20, no. 177, pp. 1–81, 2019, [Online]. Available: https://jmlr.org/papers/v20/18-760.html

[54] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," Adv. Neural Inf. Process. Syst., vol. 30, pp. 4766–4775, 2017, [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[55] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," Lancet Digit. Heal., vol. 3, no. 11, pp. e745–e750, 2021, https://doi.org/10.1016/S2589-7500(21)00208-9

[56] W. Holmes et al., "Ethics of AI in Education: Towards a Community-Wide Framework," Int. J. Artif. Intell. Educ., vol. 32, no. 3, pp. 504–526, 2022, https://doi.org/10.1007/s40593-021-00239-1

[57] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in FAT 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, 2019, pp. 59–68. https://doi.org/10.1145/3287560.3287598

[58] M. Mitchell et al., "Model cards for model reporting," in FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, 2019, pp. 220–229. https://doi.org/10.1145/3287560.3287596

[59] R. F. Wolff et al., "PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies," 2019, https://doi.org/10.7326/M18-1376

[60] R. D. Riley et al., "Calculating the sample size required for developing a clinical prediction model," BMJ, vol. 368, 2020, https://doi.org/10.1136/bmj.m441

[61] R. D. Cook, "Detection of Influential Observation in Linear Regression," Technometrics, vol. 19, no. 1, pp. 15–18, 1977, https://doi.org/10.1080/00401706.1977.10489493

[62] R. J. A. Little, "A test of missing completely at random for multivariate data with missing values," J. Am. Stat. Assoc., vol. 83, no. 404, pp. 1198–1202, 1988, https://doi.org/10.1080/01621459.1988.10478722

[63] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," J. Stat. Softw., vol. 45, no. 3, pp. 1–67, 2011, https://doi.org/10.18637/jss.v045.i03

[64] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," Stat. Med., vol. 30, no. 4, pp. 377–399, 2011, https://doi.org/10.1002/sim.4067

[65] B. Efron, "Better bootstrap confidence intervals," J. Am. Stat. Assoc., vol. 82, no. 397, pp. 171–185, 1987, https://doi.org/10.1080/01621459.1987.10478410

[66] Sture Holm, "A Simple Sequentially Rejective Multiple Test Procedure," Scand. J. Stat., vol. 6, no. 2, pp. 65–70, 1979, https://doi.org/10.2307/4615733

[67] J. G. MacKinnon and H. White, "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," J. Econom., vol. 29, no. 3, pp. 305–325, 1985, https://doi.org/10.1016/0304-4076(85)90158-7

[68] P. McCullagh, "Regression models for ordinal data," J. R. Stat. Soc. Ser. B, vol. 42, no. 2, pp. 109–127, 1980, https://doi.org/10.1111/j.2517-6161.1980.tb01109.x

[69] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995, https://doi.org/10.1007/BF00994018

[70] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015, https://doi.org/https://doi.org/10.1038/nature14539

[71] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural

networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014, [Online]. Available: https://www.jmlr.org/papers/v15/srivastava14a.html

[72] B. Efron, "Bootstrap confidence intervals for a class of parametric problems," Biometrika, vol. 72, no. 1, pp. 45–58, 1985, https://doi.org/10.1093/biomet/72.1.45

[73] R. M. O'Brien, "A caution regarding rules of thumb for variance inflation factors," Qual. Quant., vol. 41, no. 5, pp. 673–690, 2007, https://doi.org/10.1007/s11135-006-9018-6

[74] O. S. Kalange, R. S. Kahat, A. S. Kale, T. R. Kale, and P. S. Joglekar, "Implementation of Various Machine Learning Algorithms for Traffic Sign Detection and Recognition," Int. Res. J. Eng. Technol., pp. 356–361, 2022, [Online]. Available: https://www.irjet.net/volume10-issue01

[75] A. N. Muttaqin, M. Lubis, T. Mulhartono, and A. R. Lubis, "Quantifying the Causal Impact of Employment Trends on Academic Performance Using Time-Series and Public Interest Data in Indonesia," Adv. Sustain. Sci. Eng. Technol., vol. 7, no. 4, 2025, https://doi.org/10.26877/asset.v7i4.2358

[76] T. Erfando and R. Khariszma, "Sensitivity Study of the Effect Polymer Flooding Parameters To Improve Oil Recovery Using X-Gradient Boosting Algorithm," J. Appl. Eng. Technol. Sci., vol. 4, no. 2, pp. 873–884, 2023, https://doi.org/10.37385/jaets.v4i2.1871