Multitek Indonesia: Jurnal Ilmiah
Volume: 20 No. 2 December 2026, Hal. 140 - 161
ISSN: 1907-6223 (print)
ISSN: 2579-3497 (Online)
https://journal.umpo.ac.id/index.php/multitek
https://doi.org/10.24269/mtkind.v20i2.14682

Interpretable Graduate Employment Competitiveness Classification Using Catboost And Mutual Information Feature Selection

Karina Auliasari[1,2], **Syaad Patmanthara**[1], **Triyanna Widiyaningtyas**[1], **Mehaboob Mujawar**[3]
[1] Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Indonesia
[2] Department of Informatics, Institut Teknologi Nasional Malang, Indonesia
[3] Bearys Institute of Technology Mangalore, India

Corresponding Author's Email : **karina.auliasari.2605349@students.um.ac.id**

Received : September 23, 2026 **Revised : October 6, 2026** **Accepted : October 15, 2026**

Abstract

Graduate employment competitiveness is commonly predicted from academic, skill, and experiential attributes; however, severe class imbalance and data leakage may lead to misleading model performance. This study proposes an interpretable multiclass classification framework that integrates Mutual Information (MI) feature selection with CatBoost, strict data-leakage control, and imbalance-aware evaluation. The experiment used the College Graduates Employment Competitiveness dataset containing 2,000 observations classified into Low, Medium, and High employment-competitiveness classes. The identifier Student_No and the target-related Employment_Competitiveness score were excluded before model development to prevent non-substantive information and target leakage from influencing prediction. MI ranking was performed within the training process, and Top-5, Top-7, Top-9, and Top-11 feature subsets were evaluated using stratified five-fold cross-validation. CatBoost was optimized using class-balanced learning and compared with RBF-SVM and KNN. Model performance was assessed using accuracy, Macro-F1, balanced accuracy, macro precision, macro recall, and multiclass ROC-AUC, while PredictionValuesChange and permutation importance were used for interpretation. The selected Top-7 CatBoost model achieved 97.00% accuracy, 0.8798 Macro-F1, 0.8981 balanced accuracy, and 0.9936 macro ROC-AUC on the independent test set. Industry exposure was the most consistent predictive feature across MI, CatBoost importance, and permutation importance. The results demonstrate that combining MI-based feature selection, strict leakage control, and class-balanced evaluation can provide accurate and interpretable graduate employment-competitiveness classification. Nevertheless, the Low class contained only 60 observations, including 12 test cases, making its performance estimate less stable and indicating the need for validation on larger and more balanced datasets.

Keyword: CatBoost, Graduate Employment Competitiveness, Mutual Information, Feature Selection, Class Imbalance, Interpretability, Machine Learning

Copyright: ©2026 The authors. This article is published by LPPM and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

INTRODUCTION

Graduate employability has become an important concern in higher education because successful transition into employment is influenced not only by academic achievement, but also by professional skills, personal qualities, work experience, communication ability, teamwork, and the capacity to adapt to workplace demands. Recent studies emphasize that graduate employability is a multidimensional

construct that should not be reduced merely to employment status. Caballero et al. distinguish graduate employability from work readiness and explain that work readiness represents an important capability within the broader employability construct [1]. Similarly, research on workplace readiness indicates that communication competence, personal capability, technical competence, and transferable skills contribute to graduates' capacity to enter and perform effectively in the workplace [2].

The increasing availability of student and graduate data has encouraged the application of machine-learning techniques to employment and employability prediction. ElSharkawy et al. evaluated several classification algorithms, including Decision Tree, Logistic Regression, Random Forest, Naïve Bayes, and Support Vector Machine, for predicting the employability of information-technology graduates [3]. Haque et al. subsequently investigated several machine-learning classifiers using graduate tracer data and demonstrated that academic, demographic, and student-related attributes could contribute to employability prediction [4]. Andini and Lhaksmana applied Artificial Neural Network and Decision Tree models to predict student work readiness using academic knowledge, professional attitudes, soft skills, hard skills, and socioeconomic characteristics [5]. These studies demonstrate that graduate-related data contain useful predictive information; however, predictive performance alone does not necessarily reveal which attributes are used by the model, whether the selected predictors are free from target-related information leakage, or whether model performance remains reliable across imbalanced outcome classes.

More advanced approaches have recently been proposed to improve graduate-employment prediction. Using the College Graduates Employment Competitiveness dataset [6], Qi and Liu developed a reinforcement-learning framework for graduate employability prediction and career decision support, demonstrating that combinations of graduate attributes can be used to model employment competitiveness. Zheng et al. proposed an explainable hybrid deep-learning framework that integrates multi-source student information, feature selection, temporal modeling, gradient boosting, and feature-attribution techniques for multiclass employment prediction [7]. Although these approaches extend the methodological capability of employability prediction, increasingly complex architectures may also increase implementation and replication requirements. More importantly, the studies considered in the present work do not jointly address four methodological concerns that are particularly important for small tabular employability datasets: systematic feature reduction performed without information leakage, explicit control of identifier- and outcome-derived variables, evaluation that gives adequate attention to minority classes, and complementary examination of the predictors on which the final model relies. Therefore, a research gap remains for a relatively compact and reproducible tabular-learning framework that integrates these four aspects within one experimental pipeline.

Feature selection is particularly relevant in this context because graduate-employment datasets may contain variables with considerably different degrees of predictive relevance. Using all available predictors does not automatically produce a more informative model and may retain redundant, weakly informative, or conceptually inappropriate variables. Mutual Information (MI) provides a filter-based approach for quantifying statistical dependence between a predictor and a target and has been widely investigated for feature-selection problems [8], [9]. Nevertheless, MI ranking should not be treated as sufficient evidence by itself. If feature ranking is calculated using information from validation or test observations, the resulting performance estimate may become optimistic. In the present study, MI is therefore calculated only from training information, and the feature ranking is recomputed within the training folds during cross-validation. The number of selected variables is consequently determined by cross-validated Macro-F1 rather than by the final test set. This procedure allows feature reduction to be

evaluated not only in terms of dimensionality, but also in terms of its ability to preserve balanced predictive performance.

The empirical investigation uses the publicly available College Graduates Employment Competitiveness dataset [12], which contains graduate-related attributes and an Employment_Class target categorized as Low, Medium, and High. The use of these categories requires a clear conceptual limitation: the classification target represents graduate employment competitiveness as operationalized in the dataset, rather than graduate employability in its complete multidimensional meaning. Accordingly, this study does not assume that the available labels fully represent professional identity, adaptability, career sustainability, workplace readiness, or other broader aspects of employability. Instead, the dataset provides an empirical basis for examining whether observable graduate attributes can distinguish the supplied employment-competitiveness classes.

The research problem addressed in this study is therefore how to construct a compact multiclass graduate employment-competitiveness classifier that simultaneously reduces the predictor set, prevents data leakage, remains sensitive to severely imbalanced employment classes, and provides interpretable evidence about the variables used in prediction. The study specifically investigates whether an MI-selected subset of graduate attributes can retain class-balanced predictive performance when combined with CatBoost and whether the resulting predictions can be examined consistently through complementary feature-importance approaches. Accordingly, the objectives of this study are to identify a compact predictor subset using Mutual Information, prevent identifier- and target-related leakage before feature selection and model training, evaluate predictive performance using metrics appropriate for severe class imbalance, and examine the contribution of selected variables using complementary model-interpretation methods.

From an ontological perspective, the object investigated in this research is not graduate employability in its entire multidimensional reality but graduate employment competitiveness as operationally represented by observable graduate attributes and the Low, Medium, and High categories of Employment_Class available in the dataset [12]. This distinction establishes what can and cannot legitimately be claimed from the classification process. The employment-class label is treated as an empirical representation of one aspect of graduate outcomes rather than as a complete representation of employability or work readiness. Consequently, the research design defines Employment_Class as the classification target, treats substantive graduate characteristics as candidate predictors, excludes variables that do not represent meaningful graduate attributes, and avoids extending the resulting classification beyond the empirical construct represented by the dataset.

From an epistemological perspective, knowledge about graduate employment competitiveness in this study is regarded as empirical, probabilistic, and open to validation rather than as absolute knowledge about an individual graduate. Evidence is obtained through reproducible procedures consisting of MI-based feature selection, stratified data partitioning, five-fold cross-validation, CatBoost optimization, comparison with RBF-SVM and KNN, and evaluation using Macro-F1, balanced accuracy, macro precision, macro recall, and multiclass ROC-AUC. This epistemological position also requires strict separation between information used to build the model and information reserved for evaluation. Therefore, Student_No, which functions only as a record identifier, and Employment_Competitiveness, which is related to the target outcome, are excluded before feature selection and model development. The test partition is kept independent from feature ranking and hyperparameter optimization. The knowledge generated by the classifier is therefore interpreted as evidence of predictive relationships within the observed data rather than as proof of universal or causal relationships between graduate attributes and employment outcomes.

CatBoost was selected as the primary classifier because it is a gradient-boosting algorithm that constructs ensembles of symmetric decision trees and is particularly suitable for structured and tabular data [10]. Its tree-based learning mechanism can model nonlinear relationships and feature interactions while maintaining an efficient boosting structure as shown in Figure 1. In this study, CatBoost is integrated with MI-based feature selection to form the proposed MI-CatBoost framework, in which feature selection is performed before final model construction under strict training-only procedures. This combination is intended to examine whether a smaller set of empirically relevant attributes can retain competitive multiclass performance while providing a more parsimonious predictive model.

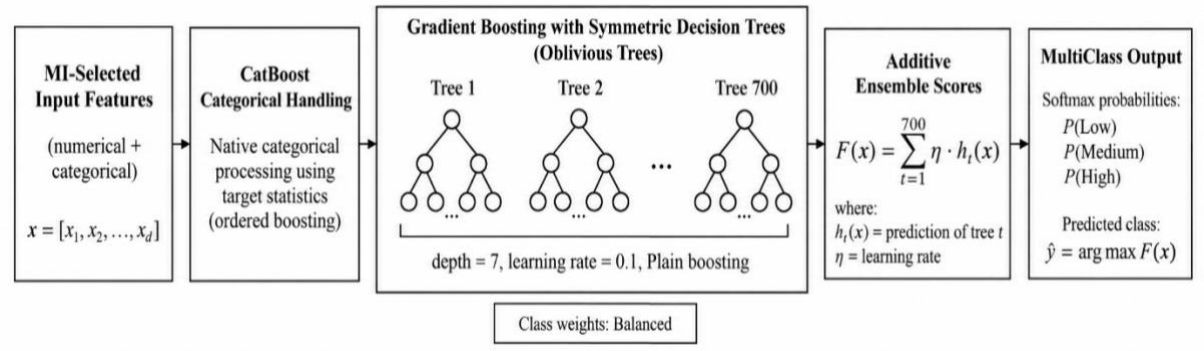


Figure 1. CatBoost Architecture[10]

From an axiological perspective, the value of the proposed framework lies not only in producing accurate classifications but also in generating results that can be examined and used responsibly by researchers and higher-education stakeholders. Predictions about graduate employment competitiveness may potentially inform employability programs, curriculum evaluation, or student career-support initiatives; therefore, model outputs should remain transparent concerning the variables on which predictions depend and should not be treated as deterministic judgments about individual graduates. Privacy, fairness, transparency, reproducibility, prevention of information leakage, and human oversight therefore constitute practical limits on the use of the proposed model. The model is intended to support analytical understanding rather than replace human academic or career-development judgment.

Model interpretation is consequently incorporated as an explicit part of the research design. Explainable and interpretable artificial-intelligence approaches have increasingly been used to examine how complex predictive models derive their outputs, including approaches that combine CatBoost with feature-attribution mechanisms [11]. In the present study, the final classifier is analyzed using two complementary global importance measures. CatBoost PredictionValuesChange represents the internal contribution assigned to each selected feature by the fitted CatBoost model, whereas Permutation Importance evaluates how model performance changes when the values of a particular predictor are randomly disturbed. Agreement or divergence between the two measures provides additional evidence about the stability of predictor relevance. Importantly, these measures are interpreted as indicators of predictive association rather than evidence of causal effects. Therefore, a feature identified as important by the model should not automatically be interpreted as an intervention that would cause improvement in graduate employment outcomes.

The main contribution of this study is an integrated MI-CatBoost framework for multiclass graduate employment-competitiveness classification that combines four methodological components within a single reproducible pipeline. First, MI-based feature ranking and cross-validated subset selection are employed to obtain a compact predictor set without consulting the independent test data. Second, explicit data-leakage control is implemented by excluding the identifier Student_No and the

target-related Employment_Competitiveness score before feature selection and model fitting. Third, the framework addresses the highly imbalanced Low, Medium, and High classes using balanced class weighting and evaluates performance through Macro-F1, balanced accuracy, class-specific metrics, and multiclass ROC-AUC in addition to overall accuracy. Fourth, the final CatBoost classifier is interpreted using both PredictionValuesChange and Permutation Importance, allowing predictive performance to be examined together with evidence regarding the variables relied upon by the model. Through these components, the proposed framework is designed to provide a compact, leakage-aware, imbalance-aware, reproducible, and interpretable approach to graduate employment-competitiveness classification.

However, the model performance is not enough when the basis of prediction is not understandable. The CatBoost classification is enriched with PredictionValuesChange and Permutation Importance to identify the features that are consistently most correlated with a model's predictions. From an axiological perspective, the value of this study lies in the provision of a more transparent and reproducible classification framework that can assist researchers and higher education stakeholders to understand the graduate attributes that are indicative of different levels of employment competitiveness, rather than solely prediction accuracy. Thus, the proposed framework includes MI-based feature ranking, training-set cross-validation for feature-subset selection, CatBoost multiclass classification, comparison with RBF-SVM and KNN, and complementary feature-importance analysis.

RESEARCH METHOD

In this study, a supervised multiclass classification framework was used to classify the graduate employment competitiveness into three levels, namely Low, Medium, and High. CatBoost was predefined as the primary classifier, while Mutual Information (MI) was used as a filter-based feature-selection method. Radial Basis Function Support Vector Machine (RBF-SVM) and K-Nearest Neighbors (KNN) were included as benchmark classifiers rather than as alternatives selected on the basis of test-set performance. Model interpretation was performed using CatBoost PredictionValuesChange and Permutation Importance. The experimental procedure consisted of eight stages: dataset inspection and descriptive analysis, variable-role definition and leakage auditing, stratified train-test partitioning, MI-based feature ranking and feature-subset selection, CatBoost hyperparameter optimization, repeated stratified stability validation, benchmark-model optimization, and final performance evaluation and model interpretation. All feature selection, preprocessing decisions, hyperparameter optimization, and validation procedures were conducted exclusively on the training partition. The independent test partition was isolated immediately after the initial split and remained unused until the final model configuration had been determined. The general workflow is illustrated in Figure 2.

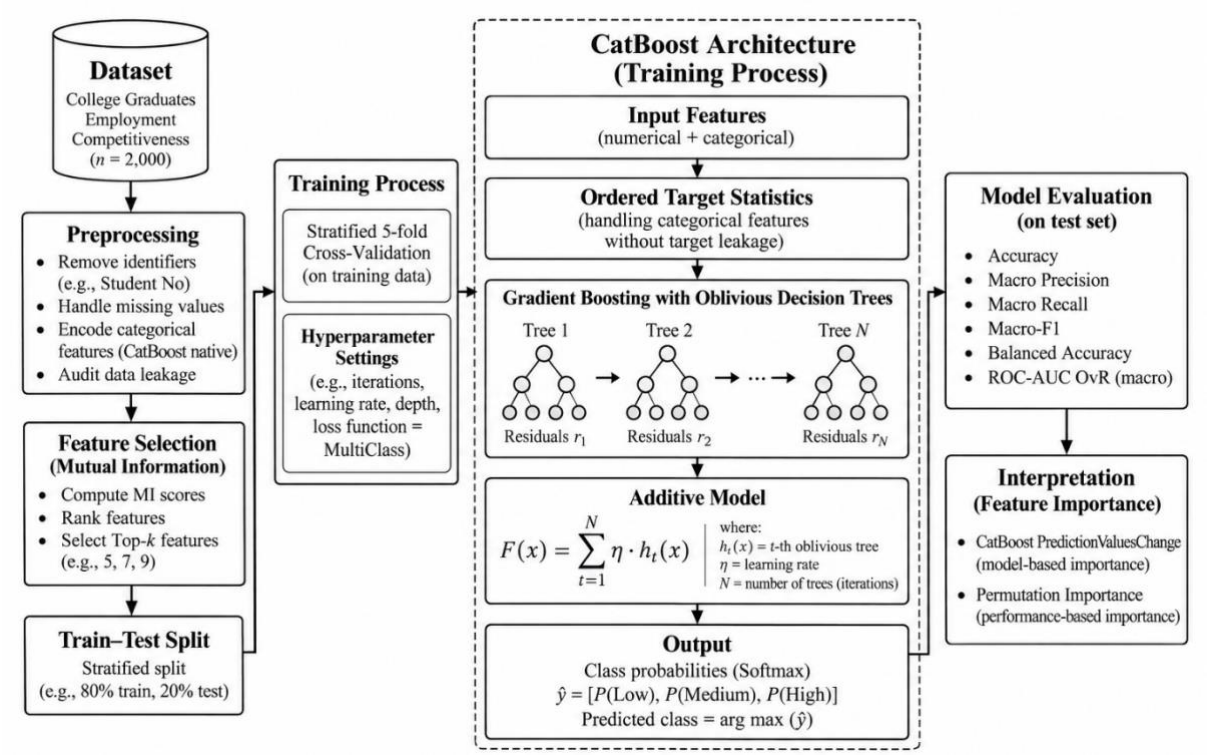


Figure 2. Research workflow for the proposed MI-CatBoost classification framework

2.1 Dataset preparation and descriptive statistics

The public dataset utilized in this study is the College Graduates Employment Competitiveness dataset from Kaggle [12]. The data set had 2,000 observations and 14 variables. - First exploration of data to understand the duplicate values, missing values, type of variables, class distribution and variables that can potentially leak target. No duplicate observations and missing values were found and all 2,000 observations were retained for further analysis.The supplied target variable, Employment_Class, was retained as the multiclass ground-truth label. The original High, Medium, and Low labels were not reconstructed from the continuous Employment_Competitiveness score. Eleven substantive variables remained as candidate predictors after leakage control: Age, Gender, Degree_Level, GPA, Programming_Skill, Communication_Skill, Teamwork, Internship_Count, Industry_Exposure, Family_Income, and Career_Motivation.The descriptive characteristics of these candidate predictors are summarized in Table 1. All predictors were stored numerically in the supplied dataset. The descriptive statistics were used solely to characterize the dataset and were not used to select features or determine the final model.

Table 1. Descriptive statistics of candidate predictors

Predictor	Mean ± SD	Median	Min–Max	Unique values
Age	24.4590 ± 2.9157	24.00	20.00–29.00	10
Gender	0.5050 ± 0.5001	1.00	0.00–1.00	2
Degree_Level	0.9980 ± 0.8212	1.00	0.00–2.00	3
GPA	7.9484 ± 1.1588	7.92	6.00–10.00	399
Programming_Skill	5.5075 ± 2.9544	6.00	1.00–10.00	10
Communication_Skill	5.5930 ± 2.8649	6.00	1.00–10.00	10
Teamwork	5.5400 ± 2.8445	6.00	1.00–10.00	10

Predictor	Mean ± SD	Median	Min–Max	Unique values
Internship_Count	1.9945 ± 1.4197	2.00	0.00–4.00	5
Industry_Exposure	4.9075 ± 3.1776	5.00	0.00–10.00	11
Family_Income	10.4890 ± 5.5249	10.47	1.01–20.00	1,240
Career_Motivation	5.4605 ± 2.8942	6.00	1.00–10.00	10

Because Gender and Degree_Level were provided as numerical codes, their numerical summaries describe the coding distribution rather than implying that these variables are continuous measurements. The target distribution was strongly imbalanced. Of the 2,000 observations, 1,791 (89.55%) belonged to the Medium class, 149 (7.45%) to the High class, and only 60 (3.00%) to the Low class. This imbalance motivated the use of stratification, balanced class weighting, Macro-F1, balanced accuracy, and class-specific performance measures rather than reliance on overall accuracy alone.

2.2 Leakage control and train-test partitioning

Two variables were excluded before any feature-ranking or model-fitting procedure. Student_No was removed because it was a unique record identifier rather than a substantive graduate attribute. Employment_Competitiveness was excluded because it represented an outcome-related continuous score associated with Employment_Class and therefore posed a direct target-leakage risk. The remaining eleven variables were eligible for MI-based feature selection.The entire data set was randomly split into training and independent testing sets, using an 80:20 stratified split with random_state = 42. This produced 1,600 training observations and 400 test observations. Stratification preserved the relative proportions of the High, Medium, and Low classes.The training partition was used for feature ranking, feature-subset selection, cross-validation, hyperparameter optimization, repeated stratified validation, and benchmark-model tuning. The 400-observation test partition was reserved exclusively for final evaluation.At no stage was the independent test partition used to calculate MI scores, determine the number or identity of selected features, optimize model hyperparameters, perform repeated validation, or select the model configuration.This separation ensured that reported hold-out performance represented evaluation on observations that had not contributed to model-development decisions.

2.3 Mutual Information Estimation and Feature-Subset Selection

Mutual Information was used to quantify the statistical dependency between each candidate predictor X and the multiclass target Y. For discrete variables, MI can be expressed in equation 1.

$$I(X;Y) = \sum_x \sum_y p(x,y)\, log \left[\frac{p(x,y)}{p(x)p(y)}\right] \quad (1)$$

Where p(x,y) denotes the joint probability distribution and p(x) and p(y) represent the corresponding marginal distributions. Predictors stored as categorical variables or having 20 or fewer unique values were treated as discrete during MI estimation. Higher-cardinality numerical predictors were treated as continuous. In the present dataset, this criterion resulted in GPA and Family_Income being handled as continuous variables, whereas predictors with lower cardinalities were handled as discrete.

For continuous predictors, MI was estimated using the non-parametric k-nearest-neighbor entropy estimation procedure implemented in mutual_info_classif, with k=3 nearest neighbors. This procedure estimates dependency from local neighbor distances rather than requiring arbitrary discretization of continuous values. The MI estimator used random_state = 42 to ensure reproducibility.

Feature selection was restricted to the training data. During selection of the number of features, MI ranking was recalculated separately within each training fold of a stratified five-fold cross-

validation. Thus, the corresponding validation fold did not participate in MI estimation or feature ranking.

Four candidate subset sizes were examined: Top-5, Top-7, Top-9, and Top-11. A fixed CatBoost configuration of 500 iterations, depth 6, learning rate 0.05, l2_leaf_reg = 5, balanced class weights, and MultiClass loss was used during this feature-subset comparison. Macro-F1 served as the selection criterion.

The Top-7 subset provided the highest mean five-fold Macro-F1 and was therefore retained. After k=7 had been determined, MI was recalculated using the complete 1,600-observation training partition to establish the final feature list. The resulting features were Industry_Exposure, Communication_Skill, Programming_Skill, GPA, Teamwork, Internship_Count, and Career_Motivation.

2.4 CatBoost Model Development and Hyperparameter Optimization

CatBoost was used as the primary multiclass classifier. Because of the severe target imbalance, auto_class_weights = "Balanced" was applied, while MultiClass was used as the optimization loss. Hyperparameter optimization was performed exclusively on the training partition using RandomizedSearchCV with 25 sampled parameter configurations and stratified five-fold cross-validation. Macro-F1 was predefined as the tuning objective. The search space consisted of 300, 500, 700, or 900 boosting iterations; tree depths from 4 to 8; learning rates of 0.03, 0.05, 0.10, or 0.15; l2_leaf_reg values of 3, 5, 7, or 9; random_strength values of 0, 0.5, 1, or 2; and border_count values of 32, 64, or 128.

The search selected 900 iterations, depth 4, learning rate 0.10, l2_leaf_reg = 7, random_strength = 2.0, and border_count = 64. The final CatBoost configuration therefore used these optimized settings together with MultiClass loss, balanced class weighting, and random seed 42. For a C-class task, model evaluation used multiclass probabilities, with the classification loss conceptually represented by cross-entropy in Equation 2.

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\, log(p_{ic}) \quad (2)$$

Where N is the number of observations, y_{ic} is the indicator that observation i belongs to class c, and p_{ic} is the predicted probability for that class.

2.5 Repeated Stratified Validation

Because the complete dataset contained only 60 Low-class observations, additional validation was performed to evaluate whether the CatBoost performance was overly dependent on a particular allocation of minority-class cases. The optimized CatBoost configuration was therefore evaluated using Repeated Stratified 5-Fold Cross-Validation with 10 repetitions, generating 50 validation runs on the training partition. A fixed random_state = 42 was used for reproducible fold generation.

For every repeated-validation split, MI ranking was recalculated exclusively from the corresponding fold-training subset. The highest-ranked seven features were then selected within that fold, preprocessing parameters were obtained from the fold-training data, and the fitted classifier was evaluated on the corresponding validation fold. Consequently, validation observations did not participate in MI estimation or preprocessing parameter estimation. The repeated procedure recorded accuracy, macro precision, macro recall, Macro-F1, balanced accuracy, and the precision, recall, and F1-score of the Low class. Because each split preserved the training class distribution, approximately 9–10 Low observations were present in each validation fold. The repeated validation was used as a stability analysis of the finalized CatBoost configuration and did not replace the independent 20% hold-out test set or introduce the test observations into model development.

2.6 Benchmark Models and Optimized Hyperparameters

RBF-SVM and KNN were used as benchmark classifiers and were trained using the same final MI-selected Top-7 feature subset as CatBoost. Since the selected features were numerically stored in the supplied dataset, the benchmark preprocessing pipeline applied median imputation where required followed by standardization. Although the dataset contained no missing observations, imputation remained part of the reproducible pipeline.

For RBF-SVM, the grid search evaluated:

$$C \in \{0.5,1,2,5,10\}$$

and

$$\gamma \in \{\text{scale}, 0.01,0.1,1.0\}$$

The RBF kernel, balanced class weighting, and probability estimation were used. Five-fold stratified cross-validation with Macro-F1 selected C=10 and γ =0.1 as the optimized hyperparameters. For KNN, the search evaluated n_neighbors $\in$ {3,5,7,9,11,15}, uniform versus distance-based weighting, and Manhattan (p=1) versus Euclidean (p=2) distance. Five-fold cross-validation with Macro-F1 selected 7 neighbors, distance weighting, and (p=2), corresponding to Euclidean distance. The optimized benchmark configurations are summarized in Table 2.

Table 2. Optimized benchmark-model hyperparameters

Model	Hyperparameter	Optimized value
RBF-SVM	Kernel	RBF
RBF-SVM	C	**10**
RBF-SVM	Gamma (γ)	**0.1**
RBF-SVM	Class weight	Balanced
KNN	Number of neighbors	**7**
KNN	Weighting	**Distance**
KNN	p	**2**
KNN	Distance metric	**Euclidean**

. The corresponding best training cross-validation Macro-F1 values were 0.9211 for RBF-SVM and 0.7486 for KNN. These values were obtained without accessing the independent test partition.

2.7 Performance Metrics and Model Interpretation

Because the Medium class accounted for 89.55% of the dataset, evaluation emphasized metrics that assign appropriate importance to minority-class performance. The reported metrics were accuracy,

macro precision, macro recall, Macro-F1, balanced accuracy, class-specific precision, recall and F1-score, the confusion matrix, and multiclass one-vs-rest macro ROC-AUC.

Macro-F1 was predefined as the principal model-selection and tuning criterion and was calculated as in Equation 3.

$$\boldsymbol{Macro\text{-}F1} = \frac{\mathbf{1}}{\boldsymbol{C}} \sum_{\boldsymbol{c=1}}^{\boldsymbol{C}} \boldsymbol{F1_c} \qquad (3)$$

Where

$$\boldsymbol{F1_c} = \frac{\mathbf{2}(\boldsymbol{Precision_c} \times \boldsymbol{Recall_c})}{\boldsymbol{Precision_c} + \boldsymbol{Recall_c}} \quad (4)$$

Macro averaging gives the same contribution to each class regardless of its frequency and was therefore appropriate for the strongly imbalanced target. After all feature-selection and tuning decisions had been finalized, the CatBoost, RBF-SVM, and KNN configurations were evaluated on the same independent 400-observation test partition. CatBoost had been defined a priori as the primary model; therefore, the hold-out results of the three algorithms were used for final comparative evaluation rather than for selecting or retuning the classifier.

The final CatBoost classifier was interpreted using two complementary global feature-importance approaches. PredictionValuesChange quantified CatBoost's internal feature contributions, while Permutation Importance measured the reduction in hold-out Macro-F1 when individual features were randomly permuted. Permutation Importance used 20 repetitions with random_state = 42 and was summarized using the mean change and its standard deviation. These importance estimates were interpreted as predictive associations and not as evidence of causal effects on graduate employment outcomes.

RESULTS AND DISCUSSION

3.1 Data quality and target imbalance

The corrected rerun confirmed that all 2,000 observations were retained after cleaning: no duplicate records were removed and no missing values were detected. The analysis therefore did not require imputation in practice, even though the modeling pipeline retained fold-wise imputation logic for reproducibility. The most important correction was methodological rather than numerical: Student_No and Employment_Competitiveness were excluded before feature selection and model fitting. Figure 3 shows a pronounced imbalance in the target, with the Medium class accounting for 89.55% of all samples. The High and Low classes accounted for only 7.45% and 3.00%, respectively. Under this distribution, a classifier could obtain a high overall accuracy by favoring Medium observations while still performing poorly for the smaller classes. Consequently, the central performance indicators in this study are Macro-F1 and balanced accuracy, supported by class-level recall and the confusion matrix.

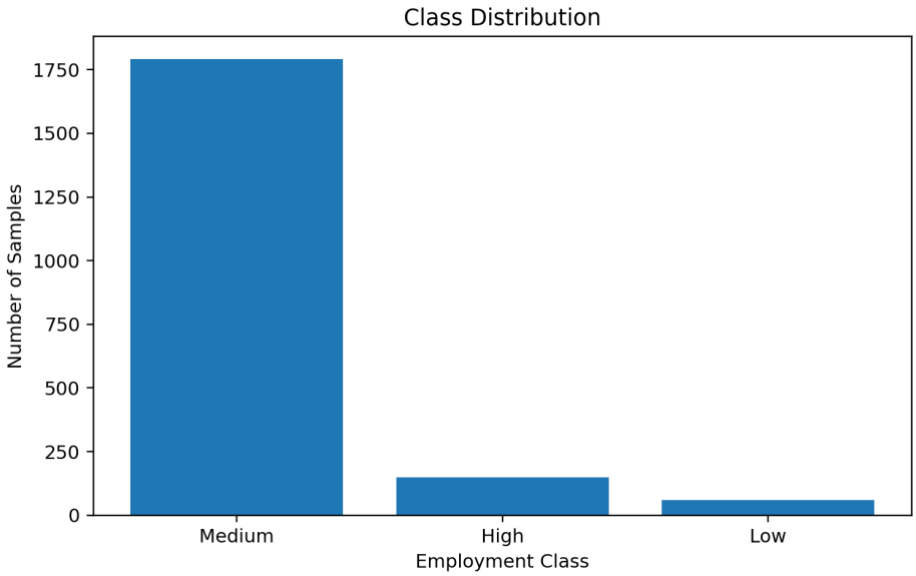


Figure 3. Distribution of High, Medium, and Low employment-competitiveness classes

3.2 Mutual Information ranking and feature-subset selection

Figure 4 and Table 3 presents the Mutual Information ranking calculated exclusively from the training partition. Industry_Exposure obtained the highest MI score (0.06591), followed by Communication_Skill (0.05590), Programming_Skill (0.04697), GPA (0.04656), and Teamwork (0.04358). A substantial decrease was observed after the five highest-ranked variables, with Internship_Count reaching 0.01125 and Career_Motivation 0.00638, whereas Age, Degree_Level, Family_Income, and Gender each produced MI scores below 0.0033. The pattern shown in Figure 4 therefore identifies a compact group of experiential, interpersonal, technical, and academic variables carrying the strongest statistical dependency with the supplied Employment_Class label. These MI values represent predictive statistical dependence and should not be interpreted as evidence of causal influence on graduate employment competitiveness.

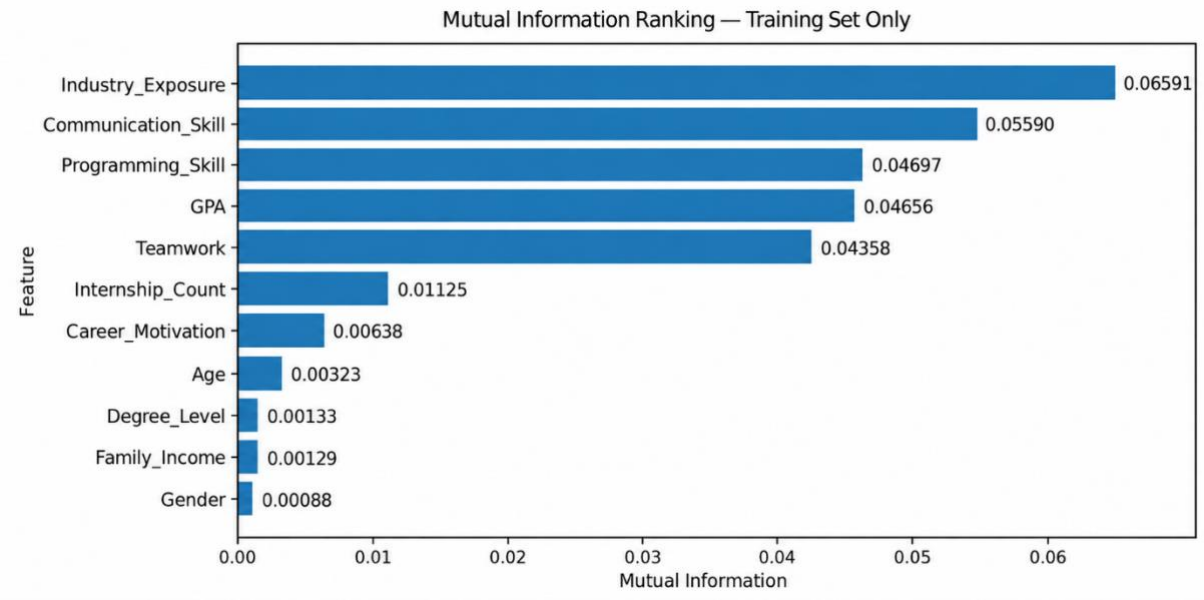


Figure 4. Mutual Information ranking calculated from the training partition only

Table 3. Mutual Information ranking on the training partition

Rank	Feature	MI score
1	Industry_Exposure	0.06591
2	Communication_Skill	0.05590
3	Programming_Skill	0.04697
4	GPA	0.04656
5	Teamwork	0.04358
6	Internship_Count	0.01125
7	Career_Motivation	0.00638
8	Age	0.00323
9	Degree_Level	0.00133
10	Family_Income	0.00129
11	Gender	0.00088

The Top-k ablation in Table 4 shows that Top-7 produced the highest mean five-fold Macro-F1, 0.8481 ± 0.0556. Top-11 (0.8470 ± 0.0365) and Top-9 (0.8466 ± 0.0512) were very close, whereas Top-5 fell to 0.7983 ± 0.0091. The mean difference between Top-7 and Top-11 was only 0.0011, but Top-7 achieved the highest cross-validation mean with four fewer variables and was therefore selected as the more parsimonious configuration under the predefined criterion.

Table 4. Feature-subset selection using stratified five-fold cross-validation

Feature subset	Mean Macro-F1	SD
Top-5	0.7983	0.0091
Top-7	0.8481	0.0556
Top-9	0.8466	0.0512
Top-11	0.8470	0.0365

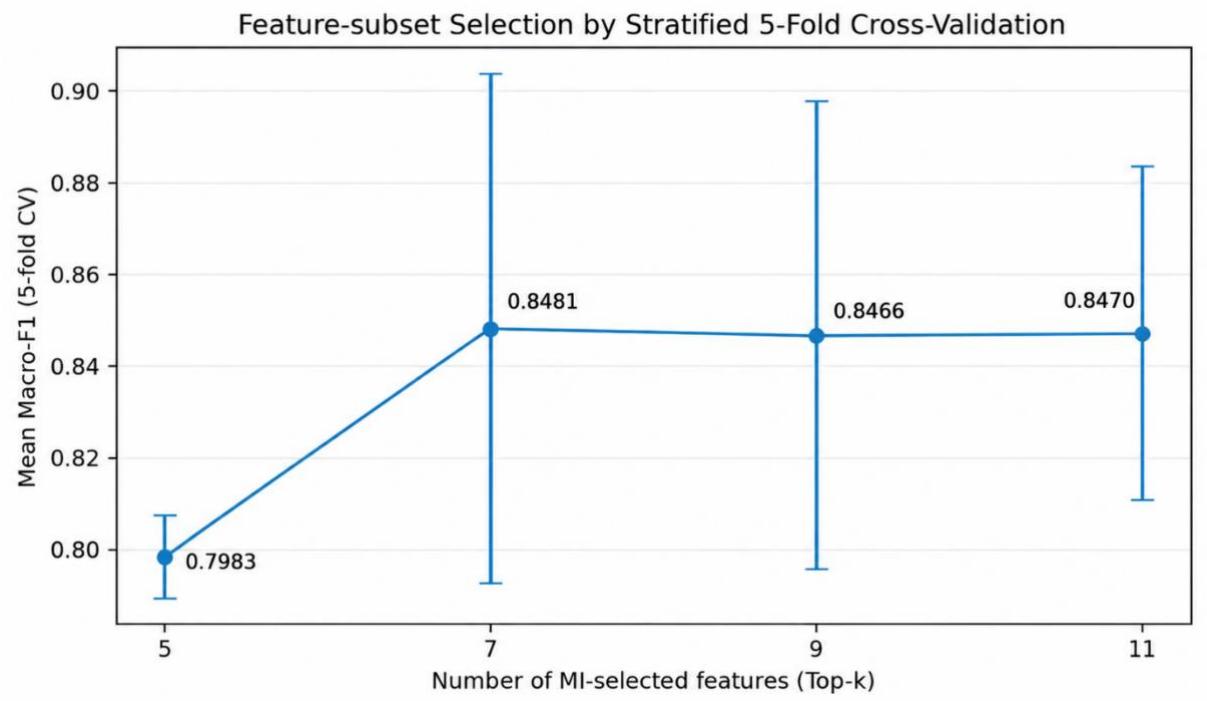


Figure 5. Mean Macro-F1 and standard deviation for the four MI-selected subset sizes

The feature-subset comparison is summarized numerically in Table 4 and visualized in Figure 5. The Top-7 configuration achieved the highest mean five-fold Macro-F1 of 0.8481 ± 0.0556, followed closely by Top-11 at 0.8470 ± 0.0365 and Top-9 at 0.8466 ± 0.0512, whereas Top-5 decreased to 0.7983 ± 0.0091. As illustrated by the overlapping variability ranges in Figure 5, the predictive differences among Top-7, Top-9, and Top-11 were small. In particular, the mean Macro-F1 difference between Top-7 and Top-11 was only 0.0011. A two-sided paired Wilcoxon signed-rank test using the five matched cross-validation folds found no statistically significant difference between these two configurations ($W = 7.0$, $p = 1.000$). Therefore, Top-7 is not claimed to be statistically superior to Top-11; instead, it was retained under the predefined selection criterion because it achieved the highest observed mean Macro-F1 while requiring four fewer predictors. Given the limited number of paired folds, this significance test should be regarded as exploratory, and repeated nested comparisons would provide stronger inferential evidence.

3.3 CatBoost optimization and final hold-out performance

After Top-7 selection, the 25-configuration randomized search produced a best cross-validation Macro-F1 of 0.8816. The selected CatBoost configuration used 900 iterations, depth 4, learning rate 0.10, l2_leaf_reg 7, random_strength 2.0, and border_count 64. The optimized parameters are summarized in Table 5.

Table 5. Optimized CatBoost hyperparameters

Hyperparameter	Selected value
random_strength	2
learning_rate	0.1
l2_leaf_reg	7
iterations	900
depth	4
border_count	64
loss_function	MultiClass
class weighting	Balanced
random seed	42

On the untouched 400-observation test partition, the final CatBoost model as shown in Table 6 achieved 97.00% accuracy, macro precision of 0.8639, macro recall of 0.8981, Macro-F1 of 0.8798, balanced accuracy of 0.8981, and macro one-vs-rest ROC-AUC of 0.9936. The test Macro-F1 was only 0.0018 lower than the best tuning cross-validation score (0.8816), indicating that the optimized model retained essentially the same class-balanced F1 performance on the held-out data rather than showing a large validation-to-test drop.

Table 6. Final CatBoost performance on the hold-out test set

Metric	**Value**
Accuracy	0.9700
Macro precision	0.8639
Macro recall	0.8981
Macro-F1	0.8798
Balanced accuracy	0.8981
ROC-AUC (OvR, macro)	0.9936
Number of selected features	7

3.4 Comparison with RBF-SVM and KNN

Table 7 and Figure 6 compare CatBoost, RBF-SVM, and KNN using the same Top-7 feature representation and the same independent hold-out test set. RBF-SVM achieved the highest accuracy (0.9750), macro precision (0.8829), and macro one-vs-rest ROC-AUC (0.9951), whereas CatBoost achieved the highest Macro-F1 (0.8798) and macro recall/balanced accuracy (0.8981). Figure 6 visually emphasizes this metric-dependent pattern: CatBoost and RBF-SVM exhibit very similar overall performance, but each model leads under different evaluation criteria. Consequently, the results do not support a claim of universal superiority for either classifier. CatBoost was retained as the primary interpretable model because Macro-F1 was the predefined optimization criterion and because the class distribution was strongly imbalanced. KNN achieved an accuracy of 0.9275 but substantially lower Macro-F1 (0.6832) and balanced accuracy (0.6018), further demonstrating why accuracy alone is insufficient for evaluating this dataset.

Table 7. Hold-out comparison of CatBoost, RBF-SVM, and KNN

Model	Acc.	Macro P	Macro R	Macro F1	Bal. Acc.	AUC
CatBoost Top-7	0.9700	0.8639	0.8981	0.8798	0.8981	0.9936
RBF-SVM	0.9750	0.8829	0.8731	0.8775	0.8731	0.9951
KNN	0.9275	0.8558	0.6018	0.6832	0.6018	0.9214

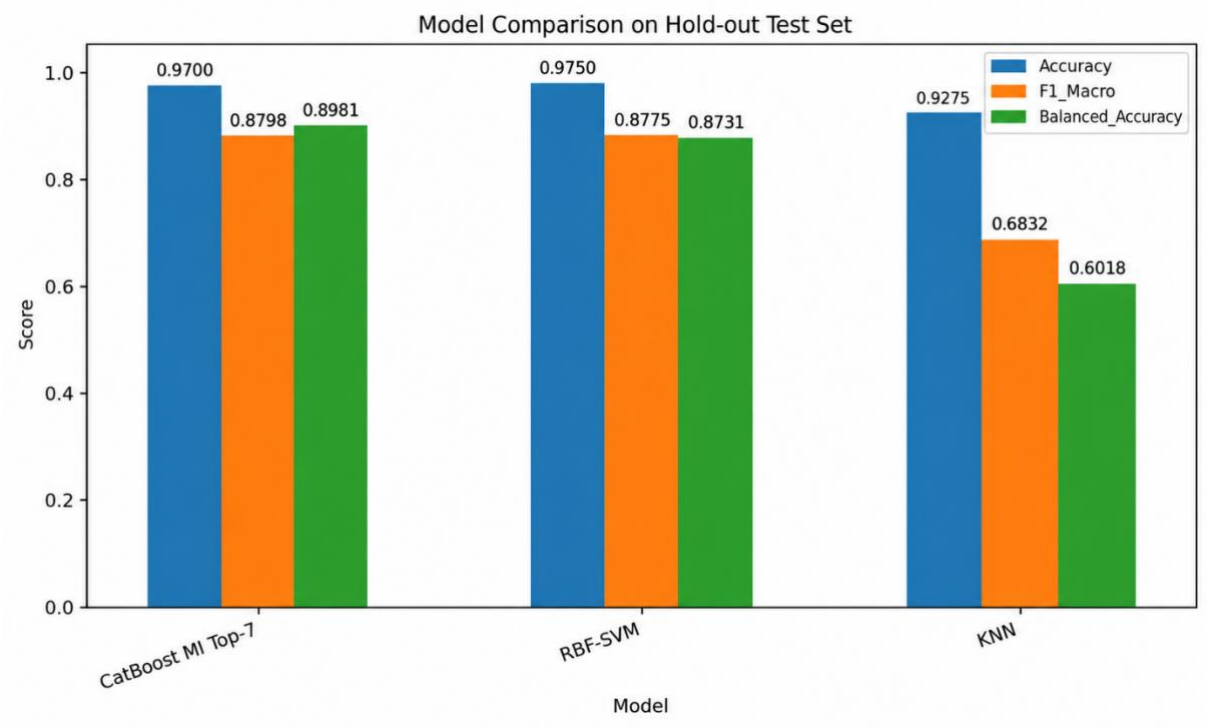


Figure 6. Comparison of accuracy, Macro-F1, and balanced accuracy on the hold-out test set

KNN achieved 0.9275 accuracy, but its Macro-F1 (0.6832) and balanced accuracy (0.6018) were substantially lower. This gap illustrates why overall accuracy alone is insufficient for this dataset: a model can remain above 92% accuracy while providing much weaker average performance across the three classes. The result also supports the use of class-balanced evaluation metrics during model selection.

3.5 Class-wise performance and error analysis

The class-wise metrics in Table 8 are consistent with the error distribution shown in Figure 7. The High class achieved a recall of 0.9667 and an F1-score of 0.9063, while the Medium class achieved precision of 0.9887, recall of 0.9777, and an F1-score of 0.9831. The Low class remained the most difficult category, with precision, recall, and F1-score all equal to 0.7500. Because only 12 Low observations were available in the hold-out test partition, these class-specific estimates should be interpreted with caution. The contrast between the weighted-average F1-score of 0.9704 and Macro-F1 of 0.8798 further demonstrates the effect of class imbalance and supports the use of macro-averaged metrics for model assessment.

Table 8. Class-specific CatBoost performance

Class	Precision	Recall	F1-score	Support
High	0.8529	0.9667	0.9062	30
Low	0.7500	0.7500	0.7500	12
Medium	0.9887	0.9777	0.9831	358
Macro average	0.8639	0.8981	0.8798	400
Weighted average	0.9714	0.9700	0.9704	400

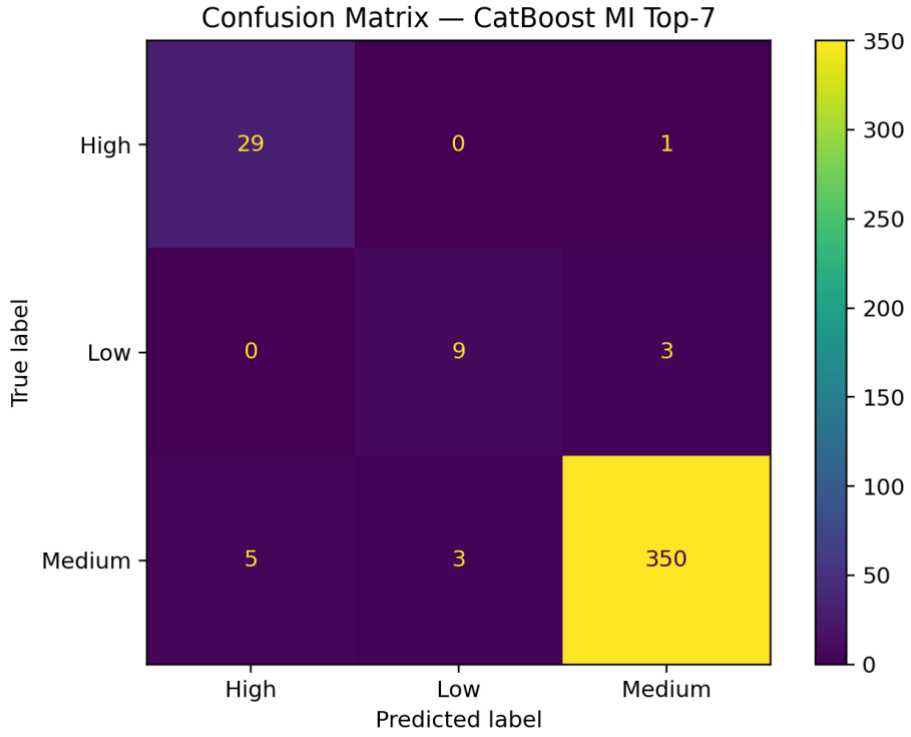


Figure 7. Confusion matrix of the final CatBoost MI Top-7 model

Figure 7 presents the confusion matrix of the final CatBoost MI Top-7 model and provides a direct view of its class-specific error pattern. CatBoost correctly classified 29 of 30 High observations, 9 of 12 Low observations, and 350 of 358 Medium observations. The 12 misclassified observations consisted of one High case predicted as Medium, three Low cases predicted as Medium, five Medium cases predicted as High, and three Medium cases predicted as Low. Importantly, Figure 7 shows no direct confusion between the High and Low classes; all errors involving these minority classes were associated with the dominant Medium category. This pattern indicates that the principal classification

challenge lies at the boundaries between each minority class and the Medium class rather than between the two minority classes themselves.

3.6 Repeated stratified validation and confidence intervals

To assess stability beyond the single 80:20 hold-out split, the optimized CatBoost pipeline was evaluated using repeated stratified 5-fold cross-validation with 10 repetitions, yielding 50 evaluation folds. As summarized in Table 9, the mean Macro-F1 was 0.8778 (SD = 0.0382; 95% CI: 0.8672–0.8884), closely matching the independent hold-out Macro-F1 of 0.8798. Mean balanced accuracy was 0.9111 (95% CI: 0.8993–0.9228), while mean accuracy reached 0.9649 (95% CI: 0.9616–0.9681). For the Low class, Table 9 also shows a repeated-validation mean F1-score of 0.7867 (SD = 0.0853; 95% CI: 0.7631–0.8104), with mean precision of 0.7512 and mean recall of 0.8453. The wider variability of the Low-class metrics relative to the overall macro metrics confirms that the smallest class remains the principal source of predictive uncertainty. In addition to the performance estimates reported in Table 9, feature-selection stability across the 50 folds showed that Industry_Exposure, Communication_Skill, GPA, Programming_Skill, and Teamwork were selected in 100% of the folds, while Internship_Count and Career_Motivation were selected in 98% and 90%, respectively, indicating substantial stability of the final Top-7 feature structure.

Table 9. Repeated stratified 5-fold cross-validation with 10 repetitions (50 evaluations)

Metric	Mean	SD	95% CI	N
Accuracy	0.9649	0.0118	0.9616-0.9681	50
Macro precision	0.8555	0.0529	0.8408-0.8701	50
Macro recall	0.9111	0.0423	0.8993-0.9228	50
Macro-F1	0.8778	0.0382	0.8672-0.8884	50
Balanced accuracy	0.9111	0.0423	0.8993-0.9228	50
Low precision	0.7512	0.1315	0.7147-0.7876	50
Low recall	0.8453	0.1029	0.8168-0.8739	50
Low F1-score	0.7867	0.0853	0.7631-0.8104	50

3.7 Interpretation of graduate employment-competitiveness predictors

CatBoost PredictionValuesChange and permutation analysis converged on Industry_Exposure as the feature with the largest predictive contribution in the final model. As summarized in Table 10, Industry_Exposure obtained a CatBoost PredictionValuesChange importance of 23.91, while permuting this feature reduced the hold-out Macro-F1 by an average of 0.3292 ± 0.0455, representing the largest performance reduction among all selected variables. Table 10 also shows that Programming_Skill, GPA, Teamwork, and Communication_Skill produced substantial predictive contributions under both importance approaches, although their relative rankings differed between the CatBoost internal importance and permutation-based evaluation. Internship_Count showed a smaller positive permutation effect (0.0869 ± 0.0276), whereas Career_Motivation had a small CatBoost importance of 3.12 but a slightly negative permutation importance of -0.0155 ± 0.0112. The agreement across the initial MI filter, the fitted CatBoost structure, and the hold-out permutation analysis identifies Industry_Exposure as the most stable predictive signal in this experiment. However, the differences between the two importance measures reported in Table 10 also indicate that model-internal importance should be interpreted together with perturbation-based evidence. These quantities describe the predictive contribution of each

variable within the trained model and should not be interpreted as evidence that modifying any of these graduate attributes would causally change employment competitiveness.

Table 10. Comparison of CatBoost and permutation importance

Feature	CatBoost importance	Permutation mean	Permutation SD
Industry_Exposure	23.91	0.3292	0.0455
Programming_Skill	15.76	0.2677	0.0341
GPA	14.11	0.2347	0.0425
Teamwork	17.32	0.2252	0.0334
Communication_Skill	18.42	0.2235	0.0356
Internship_Count	7.37	0.0869	0.0276
Career_Motivation	3.12	-0.0155	0.0112

As shown in Figure 8, the CatBoost PredictionValuesChange analysis indicates that the final model distributes its predictive contribution across several selected features. Industry_Exposure exhibited the highest model-internal importance (23.91), followed by Communication_Skill (18.42), Teamwork (17.32), Programming_Skill (15.76), GPA (14.11), Internship_Count (7.37), and Career_Motivation (3.12). Figure 9 provides a complementary permutation-based assessment and similarly identifies Industry_Exposure as the most influential predictive feature, with a mean decrease in hold-out Macro-F1 of 0.3292 ± 0.0455 when permuted. Programming_Skill ranked second in permutation importance (0.2677 ± 0.0341), followed by GPA (0.2347 ± 0.0425), Teamwork (0.2252 ± 0.0334), and Communication_Skill (0.2235 ± 0.0356). Together, Figures 8 and 9 therefore show that Industry_Exposure provides the strongest and most consistent predictive contribution, while Programming_Skill, GPA, Teamwork, and Communication_Skill also contribute substantially, although their relative rankings differ between the model-internal and perturbation-based importance measures. These five variables also formed the stable upper portion of the MI ranking, suggesting that the final classifier does not rely on a single type of information; rather, employment competitiveness is predicted from a combination of industry exposure, technical capability, academic achievement, teamwork, and communication skill.

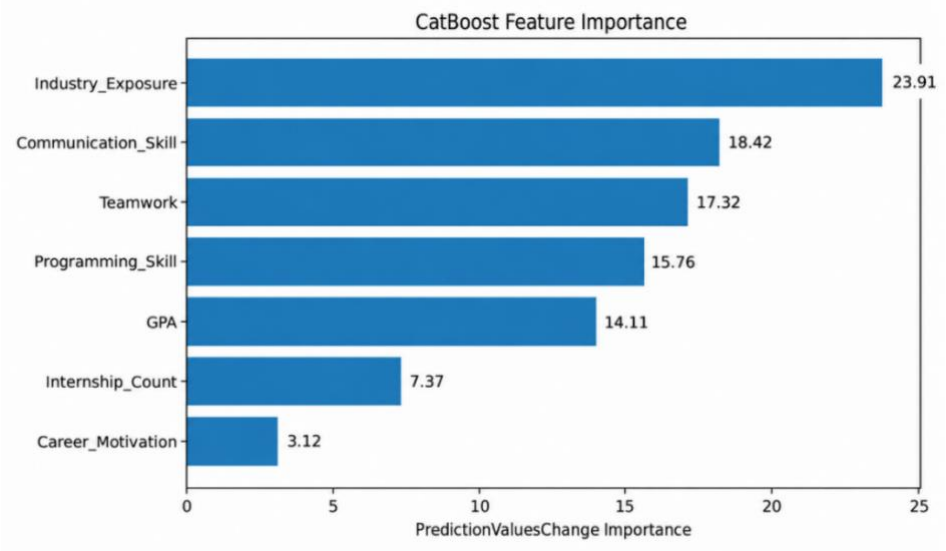


Figure 8. CatBoost PredictionValuesChange feature importance for the final Top-7 model

The contrast between Figures 8 and 9 is particularly evident for Internship_Count and Career_Motivation. Internship_Count had a moderate CatBoost PredictionValuesChange importance

of 7.37 in Figure 8 and a smaller but still positive permutation importance of 0.0869 ± 0.0276 in Figure 9, indicating that the feature provides additional predictive information but contributes less strongly than the five leading predictors. Career_Motivation showed a more pronounced divergence: as illustrated in Figure 9, CatBoost assigned it a small model-internal importance of 3.12, whereas Figure 9 shows a slightly negative permutation importance of 0.0155 ± 0.0112. A negative permutation value indicates that randomly disturbing Career_Motivation did not reduce the hold-out Macro-F1 and, on average, produced a slight increase. This finding does not imply that career motivation is harmful or unimportant in a causal sense. Rather, it suggests that, once the other six selected predictors are considered, Career_Motivation provides little robust incremental predictive information in this particular hold-out sample. The divergence between Figures 8 and 9 therefore illustrates why model-internal importance should be interpreted together with an external perturbation-based measure. Variables excluded by the Top-7 procedureAge, Degree_Level, Family_Income, and Gender had very low MI scores in the preceding feature-selection analysis and therefore were not included in the final CatBoost model or the importance analyses presented in Figures 8 and 9. Their absence from these figures should be interpreted only as evidence that they provided limited predictive information under the present dataset, MI-based selection procedure, and model configuration. It does not establish that these variables are theoretically unimportant or causally unrelated to graduate employment outcomes in other populations, institutional settings, or research designs. Accordingly, the feature-importance patterns presented in Figures 8 and 9 should be interpreted as predictive contributions within the fitted model rather than evidence of causal relationships.

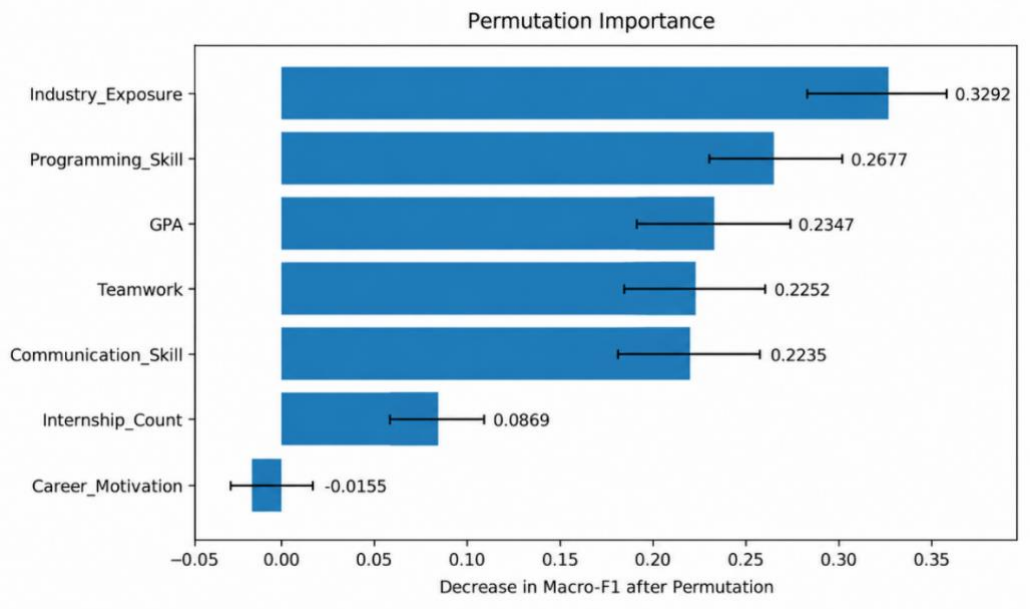


Figure 9. Permutation importance measured as the decrease in hold-out Macro-F1 over 20 repetitions

3.8 Computational behavior and reproducibility

The CatBoost randomized search evaluated 25 configurations under five-fold cross-validation and required 106.07 seconds in the recorded Google Colab execution. After hyperparameters were selected, fitting the final CatBoost model as shown in Table 11 required 1.0413 seconds and prediction for the 400-case test partition required approximately 0.0079 seconds. These timings document the recorded execution but should not be interpreted as hardware-independent benchmarks because runtime depends on the Colab environment and available resources.

Table 11. Recorded runtime of the CatBoost stages

Stage	Time (s)
CatBoost_RandomizedSearchCV_25x5	106.0737
Final_CatBoost_Fit	1.0413
Final_CatBoost_Predict	0.0079

3.8 Discussion of robustness and limitations

The corrected experiment addresses two important threats to validity in the earlier pipeline: identifier leakage and use of a target-related continuous score. It also separates the test set from feature selection and hyperparameter optimization. Recomputing MI inside each cross-validation fold is important because feature selection performed once before cross-validation can make validation performance optimistic. The tuned training CV Macro-F1 (0.8816), repeated stratified CV mean Macro-F1 (0.8778; 95% CI: 0.8672-0.8884), and independent hold-out Macro-F1 (0.8798) are closely aligned, supporting stable class-balanced performance within this dataset. However, the paired fold-wise comparison between Top-7 and Top-11 was not statistically significant (Wilcoxon $W = 7.0$, $p = 1.000$), so the Top-7 choice is justified as a compact configuration under the predefined selection rule rather than as evidence of superior predictive performance.

Nevertheless, several limitations remain. First, the class distribution is extremely imbalanced, and the Low class contains only 60 observations in total; its hold-out estimate is based on only 12 cases and its repeated-validation F1-score shows greater variability (SD = 0.0853) than the overall Macro-F1 (SD = 0.0382). Second, the target labels were accepted as supplied in Employment_Class. Because the rerun outputs do not contain the original rule used to construct High, Medium, and Low, the analysis evaluates classification of the provided labels rather than validating their underlying threshold definition. Third, all experiments use a single public dataset. Although repeated stratified validation was added and the independent 20% test set was kept untouched, external validation on an independent graduate cohort is still required before claiming transportability to other institutions, disciplines, or labor-market contexts.

Finally, the interpretability analysis is global and associative. Feature importance identifies the predictive contribution of variables used by the trained model, but it does not establish causal mechanisms. For example, the prominence of Industry_Exposure should be read as a strong predictive relationship in the observed data, not as proof that increasing industry exposure alone will cause a graduate to move to a higher employment-competitiveness class. Future work can strengthen the evidence through repeated nested comparisons of feature-subset sizes, bootstrap confidence intervals for all class-level metrics, local explanation methods, and external validation while retaining strict leakage control.

CONCLUSION

This work proposed an interpretable multiclass classification framework for the competitiveness of graduate employment by combining the Mutual Information (MI)-based feature selection method and the CatBoost model with the strict leakage control, imbalance-aware evaluation and complementary model-interpretation techniques. Student_No was dropped as a non-substantive identifier and Employment_Competitiveness was dropped due to its likely target-leakage relationship with Employment_Class. The MI-based procedure identified seven predictors Industry_Exposure, Communication_Skill, Programming_Skill, GPA, Teamwork, Internship_Count, and Career_Motivation as the final feature subset. Although Top-7 achieved the highest mean five-fold Macro-F1 among the evaluated subsets, its advantage over Top-11 was only 0.0011 and was not statistically significant in the paired Wilcoxon test ($W = 7.0$, $p = 1.000$). Therefore, Top-7 should be

interpreted as a more parsimonious configuration under the predefined selection criterion rather than as a statistically superior feature subset.On the independent hold-out test set, the optimized CatBoost model achieved an accuracy of 97.00%, Macro-F1 of 0.8798, balanced accuracy of 0.8981, and macro one-vs-rest ROC-AUC of 0.9936. Repeated stratified 5-fold cross-validation with 10 repetitions further produced a mean Macro-F1 of 0.8778 (95% CI: 0.8672–0.8884), supporting the stability of the observed predictive performance within the available dataset. The comparison with benchmark classifiers also showed that model performance depended on the evaluation criterion: RBF-SVM achieved slightly higher accuracy (97.50%) and macro ROC-AUC (0.9951), whereas CatBoost achieved higher Macro-F1 and balanced accuracy. Accordingly, the findings do not support universal superiority of CatBoost; rather, CatBoost was retained as the primary model because Macro-F1 was the predefined optimization criterion and because class-balanced performance was particularly important under the strongly imbalanced target distribution. The methodological contribution of this study lies in the combined use of MI-CatBoost feature reduction, explicit leakage control, imbalance-aware performance evaluation, repeated validation, and complementary model interpretation using CatBoost PredictionValuesChange and Permutation Importance. Industry_Exposure showed the strongest and most consistent predictive contribution across the MI ranking, CatBoost importance, and permutation analysis, while Programming_Skill, GPA, Teamwork, and Communication_Skill also contributed substantially. These findings should be interpreted as predictive associations within the fitted model rather than causal effects. In particular, the framework should not be considered a comprehensive measure of graduate employability or work readiness, because the analysis predicts the supplied Employment_Class label from the available dataset rather than measuring the full multidimensional employability construct.Several limitations remain important. The class distribution was highly imbalanced, with only 60 Low-class observations in the full dataset and 12 in the hold-out test set, making estimates for this minority class less stable. More importantly, the validity of the Employment_Class target could not be independently established because the original rules or thresholds used to construct the High, Medium, and Low categories were not revalidated in this study. Therefore, the reported results should be understood as predictive performance for the available dataset-defined employment-competitiveness labels, not as definitive evidence of graduates' overall employability. In addition, the study relied on a single public dataset, limiting generalizability across institutions, disciplines, and labor-market contexts.Future research should therefore evaluate the framework using external validation on independent graduate cohorts, as well as nested and repeated cross-validation for more robust model and feature-subset comparison. Further work should also include fairness auditing across relevant demographic or institutional subgroups, local explanation methods to support individual-level interpretation, and more systematic feature-stability analysis across repeated resampling procedures. Also, one should consider bootstrapping confidence intervals for class-specific metrics, and validating the construction of Employment_Class. These extensions would strengthen the reliability, transparency, fairness, and practical applicability of AI-assisted graduate employment-competitiveness prediction while maintaining appropriate boundaries between predictive modeling and broader interpretations of graduate employability.

REFERENCES

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

[2] S. Siddique, A. Ahsan, N. Azizi, and O. Haass, “Students’ workplace readiness: Assessment and skill-building for graduate employability,” Sustainability, vol. 14, no. 3, Art. no. 1749, 2022, doi: 10.3390/su14031749.

[3] G. ElSharkawy, Y. Helmy, and E. Yehia, “Employability prediction of information technology graduates using machine learning algorithms,” Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 10, pp. 359–367, 2022, doi: 10.14569/IJACSA.2022.0131043.

[4] R. Haque, A. Quek, C.-Y. Ting, H.-N. Goh, and M. R. Hasan, “Classification techniques using machine learning for graduate student employability predictions,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 14, no. 1, pp. 45–56, 2024, doi: 10.18517/ijaseit.14.1.19549.

[5] B. S. Andini and K. M. Lhaksmana, “Prediction of student work readiness using artificial neural network and decision tree,” Building Informatics, Technol. Sci. (BITS), vol. 7, no. 1, pp. 222–229, 2025, doi: 10.47065/bits.v7i1.7316.

[6] F. Qi and M. Liu, “An AI-driven reinforcement learning framework for graduate employability prediction and career decision support,” Discover Artif. Intell., vol. 6, Art. no. 911, 2026, doi: 10.1007/s44163-026-01612-6.

[7] H. Zheng, J. Wu, and Y. Zhang, “An explainable hybrid deep learning framework for multi-source student employment prediction,” Sci. Rep., vol. 16, Art. no. 28560, 2026, doi: 10.1038/s41598-026-59158-7.

[8] K. Robindro, U. B. Clinton, N. Hoque, and D. K. Bhattacharyya, “JoMIC: A joint MI-based filter feature selection method,” J. Comput. Math. Data Sci., vol. 6, Art. no. 100075, 2023, doi: 10.1016/j.jcmds.2023.100075.

[9] M. A. Salam, S. Basu Roy, and G. Das, “Efficient approximate top-k mutual information based feature selection,” J. Intell. Inf. Syst., vol. 61, no. 1, pp. 191–223, 2023, doi: 10.1007/s10844-022-00750-4.

[10] J. T. Hancock and T. M. Khoshgoftaar, “CatBoost for big data: An interdisciplinary review,” J. Big Data, vol. 7, Art. no. 94, 2020, doi: 10.1186/s40537-020-00369-8.

[11] M. J. D. Ebinezer and B. C. Krishna, “Explainable artificial intelligence models using SHAP enhanced CatBoost, Bi-GRU with attention, and Tab Transformer,” Sci. Rep., vol. 16, Art. no. 20357, 2026, doi: 10.1038/s41598-026-49961-7.

[12] colabsss, “College Graduates Employment Competitiveness,” Kaggle, [Online]. Available: https://www.kaggle.com/datasets/colabsss/college-graduates-employment-competitiveness. [Accessed: Sep. 19, 2026]

[13] M. A. Salam, S. Basu Roy, and G. Das, “Efficient approximate top-k mutual information based feature selection,” J. Intell. Inf. Syst., vol. 61, no. 1, pp. 191–223, 2023, doi: 10.1007/s10844-022-00750-4.

[14] H. Gong, Y. Li, J. Zhang, B. Zhang, and X. Wang, “A new filter feature selection algorithm for classification task by ensembling Pearson correlation coefficient and mutual information,” Eng. Appl. Artif. Intell., vol. 131, Art. no. 107865, 2024, doi: 10.1016/j.engappai.2024.107865.

[15] J. T. Hancock and T. M. Khoshgoftaar, “CatBoost for big data: An interdisciplinary review,” J. Big Data, vol. 7, Art. no. 94, 2020, doi: 10.1186/s40537-020-00369-8.

[16] Z. Fan, J. Gou, and S. Weng, "A feature importance-based multi-layer CatBoost for student performance prediction," IEEE Trans. Knowl. Data Eng., vol. 36, no. 11, pp. 5495–5507, 2024, doi: 10.1109/TKDE.2024.3393472.

[17] M. J. D. Ebinezer and B. C. Krishna, "Explainable artificial intelligence models using SHAP enhanced CatBoost, Bi-GRU with attention, and Tab Transformer," Sci. Rep., vol. 16, Art. no. 20357, 2026, doi: 10.1038/s41598-026-49961-7.

[18] W. Ahmed, M. A. Wani, P. Plawiak, S. Meshoul, A. Mahmoud, and M. Hammad, "Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions," Sci. Rep., vol. 15, Art. no. 26879, 2025, doi: 10.1038/s41598-025-12353-4.

[19] colabsss, "College Graduates Employment Competitiveness," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/colabsss/college-graduates-employment-competitiveness. [Accessed: Sep. 19, 2026].

[20] S. Bates, T. Hastie, and R. Tibshirani, "Cross-validation: What does it estimate and how well does it do it?," J. Am. Stat. Assoc., vol. 119, no. 546, pp. 1434–1445, 2024, doi: 10.1080/01621459.2023.2197686.