Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Exploring Machine-Learning Prediction of Self-Reported Academic Burnout

Most Afida Aktar

23320120029@stu.nsu.edu.cn

Chengdu Neusoft University https://orcid.org/0009-0007-0556-0209

Research Article

Keywords: academic burnout, higher education, machine learning, student well-being, predictive modelling, cross-validation

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11583812/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Academic burnout is a concern in higher education, but predicting students' self-reported experiences from small survey datasets presents methodological challenges. This exploratory study examined whether academic and lifestyle survey responses could distinguish three categories of students' responses to a single item asking whether they believed they were currently experiencing academic burnout: "Yes," "Not sure," and "No." The analysis used 101 survey responses and compared eight classification approaches across three feature sets. Model evaluation used repeated stratified five-fold cross-validation, with hyperparameter selection conducted within the training data. The highest mean macro-F1 score was 0.430 (SD = 0.112), achieved by the k-nearest neighbours classifier using the core feature set. Its mean accuracy was 0.485 (SD = 0.116), compared with 0.436 for the majority-class baseline. Overall performance was modest and varied across validation folds. Class-specific results showed that the model had particular difficulty identifying respondents who answered "No." Because the outcome was a single self-reported item rather than a validated burnout measure, the results should be interpreted as exploratory evidence of classification feasibility, not as a diagnostic or screening tool. The findings highlight the importance of cautious interpretation, transparent validation, and larger, independently evaluated datasets in research on student well-being.

1. Introduction

Academic burnout is an important concern in higher education because sustained academic demands may coincide with exhaustion, reduced motivation, and difficulties maintaining engagement with study. University students often manage overlapping deadlines, assessment requirements, independent learning, and responsibilities outside the classroom. Understanding how students perceive and report these experiences is relevant to research on student well-being and the conditions that support participation in higher education. However, academic burnout should not be treated as a single, easily observed condition: its interpretation depends on how it is defined and measured.

Previous research has examined burnout among university students across different populations and contexts. Schaufeli et al. (2002) investigated burnout and engagement among university students across countries. Systematic reviews have also examined the prevalence of student burnout, while reporting substantial variation across studies and measurement approaches (Kaggwa et al., 2021; Rosales-Ricardo et al., 2021). This literature establishes the importance of investigating student burnout while highlighting the need to consider differences in instruments, samples, and research designs when interpreting findings.

A response to a single survey question about burnout represents a respondent's own assessment. It should not be assumed to be equivalent to a validated multidimensional measure or a clinical assessment. This distinction matters when researchers use machine learning to classify responses. A model may learn patterns associated with how respondents answer a question, but that does not necessarily mean it measures the underlying construct accurately.

Machine learning provides computational methods for examining whether patterns across multiple variables can distinguish observed response categories. In student-related research, classification models may be useful for exploratory analysis when relationships among survey variables are complex. However, predictive performance alone does not establish that a model captures a meaningful underlying construct, identifies causal relationships, or can support decisions about individual students.

These concerns are particularly important when the available dataset is small and predictors include self-reported experiences that may overlap conceptually with the outcome. Small-sample prediction can be sensitive to the composition of the sample, the number and type of candidate predictors, and the validation procedure. Vabalas et al. (2019) discussed the risk of biased performance estimates in machine-learning studies with limited sample sizes. Nested cross-validation can help separate model selection from performance evaluation, but it cannot eliminate uncertainty caused by limited data or replace evaluation on an independent sample.

Against this background, the present study examines 101 survey responses using several machine-learning classification approaches. The outcome is a single self-reported question about academic burnout, recorded in three categories: “Yes,” “Not sure,” and “No.” The analysis compares performance across three feature sets that differ in their inclusion of questionnaire items. This comparison explores whether classification performance changes when particular groups of predictors are excluded. It does not establish the independent causal importance of any variable.

The study addresses three research questions:

1. How well do selected machine-learning classifiers distinguish among three categories of students’ self-reported responses to a single academic-burnout question?
2. How does classification performance vary across feature sets with different combinations of survey items?
3. What methodological and interpretive limitations arise when using a small, cross-sectional self-report survey to explore classification of academic-burnout responses?

The study’s contribution is exploratory. It documents the performance of several classifiers under a common internal validation procedure, examines the sensitivity of results to feature-set composition, and identifies requirements for stronger future research. These requirements include larger and more diverse samples, validated measurement instruments, reproducible evaluation, and independent validation.

The study does not establish causal relationships between academic or lifestyle factors and burnout. It also does not develop a tool suitable for diagnosing burnout or making decisions about individual students. Its findings must be interpreted in light of the outcome’s limited measurement scope, the sample size, and the absence of external validation.

2. Literature Review

2.1. Academic Burnout in Higher Education

Academic burnout is commonly discussed as a response to sustained academic demands and may involve exhaustion, disengagement from study, and reduced feelings of academic efficacy. Research on university students has examined burnout alongside academic engagement, motivation, and student well-being. Schaufeli et al. (2002) examined burnout and engagement among university students across countries, demonstrating the relevance of studying these experiences in higher education.

Systematic reviews have investigated the prevalence of burnout among university students. Kaggwa et al. (2021) reviewed evidence from university students in low- and middle-income countries, while Rosales-Ricardo et al. (2021) reviewed research on burnout prevalence among university students. Their work highlights the importance of the topic but also the need to interpret prevalence estimates in the context of different samples, instruments, and study methods.

Burnout should not be treated as interchangeable with stress, anxiety, tiredness, or poor well-being. These experiences may overlap, but they are not identical constructs. Measurement therefore matters. A study using a validated multi-item instrument assesses a different outcome from a study asking respondents whether they believe they are experiencing burnout.

This distinction is central to the present study. The outcome variable is based on one self-report question asking whether respondents believe they are currently experiencing academic burnout. Responses are grouped into “Yes,” “Not sure,” and “No.” These categories represent respondents’ perceptions rather than a validated burnout score or clinical assessment. Accordingly, this study examines the classification of self-reported responses and does not claim to diagnose burnout.

2.2. Academic and Lifestyle Factors

Academic burnout is studied alongside factors such as perceived workload, academic stress, motivation, concentration, sleep, physical activity, and psychological well-being. These variables may provide useful context for understanding students’ experiences, although their relationships with burnout can be complex. For example, poor sleep and academic stress may accompany feelings of exhaustion, but a cross-sectional survey cannot determine whether one factor causes another.

A demands-and-resources perspective provides one way to consider student experiences in relation to academic pressures and available resources. Jagodics and Szabó (2023) examined student burnout in higher education through this general perspective. Such approaches discourage the assumption that a single factor explains burnout and instead encourage attention to the context in which students experience academic demands.

The present study includes survey responses relating to academic experiences and lifestyle. Some candidate predictors may be conceptually close to the self-reported burnout outcome. This creates an interpretive challenge: a model may classify responses partly because related survey items reflect similar experiences. Consequently, even relatively strong classification performance would not, by itself, establish that a variable is an independent cause or an early warning indicator of burnout.

To examine this issue, the study compares feature sets containing different combinations of questionnaire items. The comparison is intended to explore how performance changes with predictor composition. It is not a causal analysis, and differences in performance cannot establish the independent importance of any particular factor.

2.3. Machine Learning in Student-Related Research

Machine learning includes methods that learn patterns from data and use those patterns to predict a target category. Classification algorithms such as logistic regression, support vector machines, decision trees, random forests, and k-nearest neighbours use different approaches to represent relationships between predictors and target labels.

The use of machine learning does not automatically produce a reliable or generalisable prediction tool. Performance depends on data quality, sample size, measurement, class distribution, modelling choices, and evaluation procedures. A model may perform well on a particular dataset but fail to generalise to new respondents or student populations.

This concern is especially important for small datasets. When only a limited number of observations are available, the inclusion or exclusion of a few respondents can materially affect model fitting and evaluation. A single train–test split may consequently provide an unstable estimate of performance. Vabalas et al. (2019) examined validation methods in limited-sample settings and highlighted the importance of separating model development from evaluation.

The present study uses repeated stratified five-fold cross-validation and performs hyperparameter selection within each outer training partition. This design provides multiple internal performance estimates while separating tuning from the corresponding outer test fold. Nevertheless, repeated cross-validation does not compensate for an unrepresentative sample, create new independent participants, or establish external validity. The resulting estimates remain exploratory.

2.4. Evaluating Multiclass Classification

Accuracy is a commonly reported classification metric, but it can be misleading when target classes differ in size. A model that predominantly predicts the most frequent class may achieve reasonable accuracy while performing poorly on other categories. The present study therefore evaluates accuracy alongside balanced accuracy and macro-averaged precision, recall, and F1-score.

Macro-F1 calculates the F1-score separately for each class and then averages the scores, giving each class equal weight. Balanced accuracy averages recall across classes. These measures help assess whether performance extends beyond the most frequent response category. A majority-class baseline is also useful for determining whether fitted models improve on a simple prediction strategy.

Performance averages should be interpreted alongside their variability and class-specific results. The standard deviations reported in this study describe variation across outer-fold evaluations; they are not confidence intervals. Moreover, repeated cross-validation produces overlapping training and test partitions,

so fold-level estimates are not fully independent. Confusion matrices and class-specific metrics are therefore important for identifying errors that overall scores may conceal (Sokolova & Lapalme, 2009).

2.5. Research Gap and Study Contribution

Existing research has examined academic burnout, associated student experiences, and the use of computational methods in psychological and student-related research. However, findings from studies using validated burnout instruments, larger samples, or different predictor sources cannot be directly transferred to a small survey in which the target is a single self-reported question.

The present study addresses this narrower methodological problem by comparing eight classification approaches across three feature sets using 101 survey responses. It examines whether the models can distinguish “Yes,” “Not sure,” and “No” responses and compares their performance with a majority-class baseline.

The contribution is exploratory rather than clinical or operational. The study does not claim to develop a validated screening instrument, identify causal risk factors, or demonstrate that the models are suitable for student-support decisions. The survey also did not directly measure educational-technology use or technology-supported learning. Therefore, the study cannot establish the effects of educational technology on academic burnout.

3. Methodology

3.1. Research Design

This study used an exploratory quantitative design to investigate whether machine-learning classifiers could distinguish students’ self-reported academic-burnout responses using academic and lifestyle survey data. Eight classification approaches were compared across three feature sets. The focus was predictive feasibility and methodological comparison rather than causal explanation or clinical assessment.

The analysis workflow comprised data preparation, feature-set construction, model development, nested cross-validation, and performance evaluation. Preprocessing and hyperparameter selection were incorporated into the validation procedure to reduce information leakage between training and evaluation data.

3.2. Data Source and Sample

The dataset consisted of 101 responses collected through a Google Forms questionnaire. The survey link was shared through student groups and social media. The questionnaire covered demographic characteristics, academic experiences, stress-related experiences, and lifestyle factors. Before analysis-related exclusions, the dataset contained 30 columns.

The target variable was derived from the question, “Do you believe you are currently experiencing academic burnout?” Respondents selected one of three categories: “Yes,” “Not sure,” or “No.” The observed

distribution was 44 “Yes” responses (43.56%), 33 “Not sure” responses (32.67%), and 24 “No” responses (23.76%).

Table 1
Distribution of self-reported academic-burnout responses

Response category	Frequency	Percentage
Yes	44	43.56%
Not sure	33	32.67%
No	24	23.76%
Total	**101**	**100.00%**

The target represents respondents’ own perceptions and should not be interpreted as a validated measurement of burnout or a clinical diagnosis. The sample size also limits the precision and generalisability of the model-performance estimates.

3.3. Outcome and Predictor Variables

The target variable was the three-category response to the academic-burnout question. Predictor variables were drawn from the questionnaire’s demographic, academic, stress-related, and lifestyle items.

The timestamp, consent item, optional email field, target question, open-text responses, and blank columns were excluded from the predictor sets. The optional email field was not used as a model input. Open-text responses were excluded because the analysis focused on structured survey features.

Three feature sets were defined to examine how predictor inclusion affected classification performance:

1. **Full**: Q2–Q20 and Q22–Q25, comprising 23 requested predictors.
2. **Core**: Q2–Q20, comprising 19 requested predictors.
3. **Reduced**: Q2–Q6, Q8–Q10, and Q16–Q20, comprising 13 requested predictors.

Table 2
Predictor composition of the three feature sets

Feature set	Questionnaire items	Number of predictors	Purpose
Full	Q2–Q20 and Q22–Q25	23	Evaluate the full specified structured predictor set
Core	Q2–Q20	19	Exclude Q22–Q25
Reduced	Q2–Q6, Q8–Q10, and Q16–Q20	13	Exclude Q7, Q11–Q15, and Q22–Q25

The feature-set comparison was designed as an exploratory sensitivity analysis. It can indicate whether classification performance changes when specified questionnaire items are removed. It cannot establish

that any retained item is an independent cause of burnout or that a particular feature set measures burnout more accurately.

3.4. Data Preprocessing

Preprocessing was implemented within model-training pipelines so that preprocessing parameters were estimated from training data rather than from the held-out outer test fold. This is important because preprocessing the complete dataset before cross-validation can allow information from evaluation observations to influence model development.

For numerical variables, missing values were imputed using the median, missingness indicators were added, and numerical features were standardised. For categorical variables, missing values were imputed using the most frequent category, followed by one-hot encoding.

Standardisation can be expressed as:

[

z=\frac{x-\mu_{\mathrm{train}}}{\sigma_{\mathrm{train}}}

]

where (x) is the observed value, (\mu_{\mathrm{train}}) is the mean estimated from the training partition, and (\sigma_{\mathrm{train}}) is the corresponding standard deviation. These training-derived values were applied to validation or test data without being recalculated from those partitions.

For numerical imputation, the median was estimated from the training partition. For categorical imputation, the most frequent training-partition category was used. One-hot encoding represented categorical values as binary indicator features. These operations were fitted separately within the relevant training pipeline.

3.5. Classification Approaches

Eight classification approaches were evaluated: a majority-class baseline, regularised logistic regression, radial-basis-function support vector machine (RBF SVM), linear support vector machine, random forest, Extra Trees, decision tree, and k-nearest neighbours (KNN).

The majority-class baseline predicted the most frequent class in the training data. Logistic regression provided a regularised linear classification approach. The two SVM approaches evaluated linear and nonlinear decision boundaries. Random forest and Extra Trees used ensembles of decision trees, while the single decision tree provided a tree-based comparison. KNN classified observations according to the labels of nearby observations in the feature space.

The candidate models were selected to compare different classification strategies using the same dataset and feature sets. Given the limited sample size, the purpose was not to establish that one algorithm is universally superior, but to compare observed performance under a common evaluation procedure.

Table 3
Classification approaches and candidate hyperparameters

Classifier	Candidate hyperparameters
Majority-class baseline	Predict the most frequent training class
Regularised logistic regression	(C \in {0.01, 0.1, 1, 10})
RBF SVM	(C \in {0.1, 1, 10}); (\gamma \in {\text{scale}, 0.01, 0.1})
Linear SVM	(C \in {0.01, 0.1, 1, 10})
Random forest	max_features ∈ {sqrt, 0.5}; min_samples_leaf ∈ {2, 5}; max_depth ∈ {None, 4, 8}
Extra Trees	max_features ∈ {sqrt, 0.5}; min_samples_leaf ∈ {2, 5}; max_depth ∈ {None, 4, 8}
Decision tree	min_samples_leaf ∈ {2, 5, 10}; max_depth ∈ {2, 3, None}; criterion ∈ {gini, entropy}
K-nearest neighbours	n_neighbors ∈ {3, 5, 7}; weights ∈ {uniform, distance}

The logistic regression, SVM, random forest, Extra Trees, and decision tree implementations used class weighting to address differences in class frequencies. The exact implementation, including estimator settings and the baseline procedure, should be preserved in the analysis script so that the reported results can be reproduced.

3.6. Nested Repeated Stratified Cross-Validation

Model evaluation used repeated stratified five-fold outer cross-validation with five repeats and a fixed random seed of 42. This generated 25 outer train–test partitions for each model and feature-set combination. Stratification aimed to preserve the distribution of the three target classes across folds.

Within each outer training partition, an inner cross-validation procedure selected hyperparameters from the predefined candidate grids. The inner procedure used up to three folds, with the number of folds reduced when required by the smallest class in the training data. Hyperparameter selection used macro-F1. After selection, the chosen pipeline was fitted on the outer training partition and evaluated on the corresponding outer test partition.

Nested cross-validation separates hyperparameter selection from outer performance evaluation. This helps reduce optimistic bias that can arise when the same validation results are used both to tune a model and to report its performance (Vabalas et al., 2019).

Let (D_{\mathrm{train}}^{(j)}) and (D_{\mathrm{test}}^{(j)}) represent the training and test partitions for outer fold (j). The model pipeline was tuned using only (D_{\mathrm{train}}^{(j)}). Predictions were then evaluated on (D_{\mathrm{test}}^{(j)}), which was not used to select the hyperparameters for that fold.

The mean performance across outer folds for a metric (M) is:

[

\bar{M}=\frac{1}{K}\sum_{j = 1}^{K}M_j

]

where (M_j) is the metric for outer fold (j), and (K = 25) is the number of outer-fold evaluations per model and feature set.

The sample standard deviation across outer folds is:

[

s_M=\sqrt{\frac{1}{K-1}\sum_{j = 1}^{K}(M_j-\bar{M})^2}

]

The standard deviation describes observed variability across outer-fold results. It is not a confidence interval, and the repeated-fold estimates are not fully independent because the training and test partitions overlap across repetitions.

3.7. Evaluation Metrics

Performance was assessed using accuracy, balanced accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1-score. These metrics were selected to provide a broader view of multiclass performance than accuracy alone.

For a class (c), precision is:

[

\mathrm{Precision}_c=\frac{TP_c}{TP_c + FP_c}

]

Recall is:

[

\mathrm{Recall}_c=\frac{TP_c}{TP_c + FN_c}

]

where (TP_c), (FP_c), and (FN_c) represent true positives, false positives, and false negatives for class (c). The F1-score is the harmonic mean of precision and recall:

[

F1_c = 2\cdot

\frac{\mathrm{Precision}_c\cdot\mathrm{Recall}_c}

{\mathrm{Precision}_c+\mathrm{Recall}_c}

]

Macro-averaged precision, recall, and F1 are calculated by giving each class equal weight:

[

\mathrm{MacroPrecision}=\frac{1}{C}\sum_{c = 1}^{C}\mathrm{Precision}_c

]

[

\mathrm{MacroRecall}=\frac{1}{C}\sum_{c = 1}^{C}\mathrm{Recall}_c

]

[

\mathrm{MacroF1}=\frac{1}{C}\sum_{c = 1}^{C}F1_c

]

In this study, (C = 3), corresponding to “Yes,” “Not sure,” and “No.”

Accuracy is the proportion of correctly classified observations:

[

\mathrm{Accuracy}=

\frac{\text{Number of correct predictions}}

{\text{Total number of predictions}}

]

Balanced accuracy is the mean recall across target classes:

[

\mathrm{BalancedAccuracy}=

\frac{1}{C}\sum_{c = 1}^{C}\mathrm{Recall}_c

]

Macro-F1 and balanced accuracy were included because the class distribution was unequal and the study sought to assess performance across all three response categories. Accuracy was retained for interpretability and comparison with the majority-class baseline. For each model and feature set, the mean and standard deviation of the outer-fold metrics were reported.

3.8. Implementation and Reproducibility

The analysis was implemented in Python using a scripted machine-learning workflow. A fixed random seed, predefined feature sets, specified hyperparameter grids, and consistent evaluation metrics were used to support reproducibility. Data preprocessing and hyperparameter selection were performed within the cross-validation framework to reduce information leakage between training and evaluation data.

Model performance was evaluated using repeated stratified five-fold cross-validation across the three feature sets. The implementation included systematic hyperparameter selection and consistent calculation of the evaluation metrics. Given the exploratory nature of the study and the relatively small sample size, the results should be interpreted cautiously and validated using larger, independent datasets.

3.9. Ethical considerations

This study used an online survey to collect self-reported responses from university students. All participants provided informed consent before participating in the survey.

4. Results

4.1. Overall Classification Performance

Eight classification approaches were evaluated across the Full, Core, and Reduced feature sets. Accuracy, balanced accuracy, and macro-F1 were calculated across the 25 outer-fold evaluations. The majority-class baseline achieved a mean accuracy of 0.436 and a mean macro-F1 of 0.202 across all three feature sets.

The highest mean macro-F1 was obtained by KNN using the Core feature set, with a mean macro-F1 of 0.430 (SD = 0.112), mean accuracy of 0.485 (SD = 0.116), and mean balanced accuracy of 0.443 (SD = 0.107). This configuration performed better than the majority-class baseline on these metrics. However, its mean accuracy remained below 0.50, and variation across outer folds was substantial relative to the observed performance. The results therefore indicate limited classification performance rather than a highly reliable predictive model.

KNN also achieved the highest mean macro-F1 within the Full feature set, at 0.410 (SD = 0.109), and within the Reduced feature set, at 0.386 (SD = 0.099). The Core feature set produced the strongest KNN results among the three feature configurations. This pattern suggests that feature-set composition was associated with differences in observed performance for KNN. It does not establish that the excluded variables are unhelpful in other samples or that the Core feature set is generally superior.

Table 4
Mean classification performance across outer cross-validation folds

Feature set	Model	Accuracy	Balanced accuracy	Macro-F1
Full	Majority baseline	0.436 ± 0.020	0.333 ± 0.000	0.202 ± 0.007
Full	Logistic regression	0.379 ± 0.094	0.356 ± 0.085	0.345 ± 0.084
Full	RBF SVM	0.432 ± 0.107	0.408 ± 0.105	0.384 ± 0.099
Full	Linear SVM	0.381 ± 0.118	0.347 ± 0.105	0.333 ± 0.097
Full	Random forest	0.380 ± 0.099	0.364 ± 0.098	0.351 ± 0.097
Full	Extra Trees	0.422 ± 0.107	0.401 ± 0.105	0.387 ± 0.101
Full	Decision tree	0.386 ± 0.093	0.383 ± 0.099	0.365 ± 0.100
Full	KNN	0.477 ± 0.106	0.431 ± 0.099	0.410 ± 0.109
Core	Majority baseline	0.436 ± 0.020	0.333 ± 0.000	0.202 ± 0.007
Core	Logistic regression	0.339 ± 0.111	0.323 ± 0.113	0.316 ± 0.107
Core	RBF SVM	0.416 ± 0.119	0.391 ± 0.113	0.374 ± 0.116
Core	Linear SVM	0.400 ± 0.091	0.368 ± 0.085	0.360 ± 0.089
Core	Random forest	0.422 ± 0.122	0.409 ± 0.123	0.392 ± 0.122
Core	Extra Trees	0.408 ± 0.114	0.386 ± 0.110	0.376 ± 0.108
Core	Decision tree	0.389 ± 0.123	0.382 ± 0.130	0.369 ± 0.134
Core	KNN	0.485 ± 0.116	0.443 ± 0.107	0.430 ± 0.112
Reduced	Majority baseline	0.436 ± 0.020	0.333 ± 0.000	0.202 ± 0.007
Reduced	Logistic regression	0.348 ± 0.104	0.330 ± 0.106	0.323 ± 0.101
Reduced	RBF SVM	0.389 ± 0.125	0.367 ± 0.117	0.354 ± 0.114
Reduced	Linear SVM	0.367 ± 0.096	0.329 ± 0.086	0.316 ± 0.084
Reduced	Random forest	0.378 ± 0.106	0.365 ± 0.110	0.354 ± 0.107
Reduced	Extra Trees	0.420 ± 0.108	0.395 ± 0.101	0.384 ± 0.105
Reduced	Decision tree	0.323 ± 0.083	0.327 ± 0.086	0.299 ± 0.088
Reduced	KNN	0.444 ± 0.093	0.406 ± 0.097	0.386 ± 0.099

Note. Values are means ± standard deviations across outer-fold evaluations. SD describes observed fold-to-fold variation; it is not a confidence interval. The Core feature-set name in the original analysis output, Core_no_Q22_Q25, is shortened to “Core” here.

4.2. Comparison Across Classifiers

Performance varied across classifiers and feature sets. KNN produced the highest mean macro-F1 in all three feature sets, while Extra Trees and RBF SVM also achieved comparatively higher macro-F1 values than several other approaches in the Full and Reduced configurations. The lowest mean macro-F1 in the comparison was 0.299 for the Reduced-feature decision tree.

The majority-class baseline achieved higher mean accuracy than several fitted models, including logistic regression, linear SVM, random forest, and decision tree under some feature configurations. This illustrates why accuracy should not be considered in isolation. The baseline's balanced accuracy was approximately 0.333, consistent with predicting the most frequent class without effectively distinguishing all three categories.

Several fitted models achieved higher balanced accuracy and macro-F1 than the baseline, but their performance remained modest. The difference between the highest KNN macro-F1 and the baseline macro-F1 was approximately 0.228. This is a descriptive difference between reported means and should not be interpreted as a statistically significant improvement. No statistical significance test or independent external validation is reported.

4.3. Comparison of Feature Sets

For KNN, mean macro-F1 decreased from 0.430 with the Core feature set to 0.410 with the Full feature set and 0.386 with the Reduced feature set. Mean balanced accuracy followed the same ordering: 0.443 for the Core set, 0.431 for the Full set, and 0.406 for the Reduced set.

These results indicate that the number and composition of predictors alone did not determine performance. The Reduced set contained fewer predictors but did not outperform the other configurations. Removing selected survey items may change the information available to a classifier, but the observed results do not show that one feature set is universally optimal.

The Core feature set's higher performance should be interpreted cautiously because the sample contained only 101 responses and the outcome was based on a single self-report question. Some retained variables may also overlap conceptually with the target. The comparison therefore provides exploratory evidence about model sensitivity to feature-set composition, rather than evidence of independent predictors of academic burnout.

4.4. Class-Specific Performance

Class-specific metrics provide additional detail about classification errors. For the KNN model using the Core feature set, pooled outer-test predictions produced an F1-score of 0.224 for "No," 0.515 for "Not sure," and 0.571 for "Yes." Recall was lowest for the "No" category (0.183), compared with 0.527 for "Not sure" and 0.618 for "Yes."

These results indicate that the model was substantially less effective at identifying respondents who selected "No" than those who selected "Yes." The "Not sure" category also showed moderate classification performance. Consequently, the overall model metrics conceal meaningful differences in performance across response categories.

Table 5
Class-specific performance of Core-feature KNN

Self-reported response	Precision	Recall	F1-score	Support across repeats
No	0.289	0.183	0.224	120
Not sure	0.503	0.527	0.515	165
Yes	0.531	0.618	0.571	220

Note. Metrics are calculated from pooled outer-test predictions. Each respondent contributes one prediction per repeat, giving five predictions per respondent across the five repeats. Support therefore represents repeated predictions, not the number of unique respondents.

The class-specific results reinforce the exploratory nature of the findings. In particular, the low recall for “No” means that many respondents in this category were incorrectly classified into another category. The model should not be treated as a reliable individual-level screening tool. Reporting class-specific precision, recall, and F1 alongside aggregate metrics helps reveal limitations in multiclass classification (Sokolova & Lapalme, 2009).

4.5. Confusion-Matrix Analysis

The confusion matrix for the Core-feature KNN model provides further insight into classification errors. Across the five repetitions of outer cross-validation, the model correctly classified 22 of 120 pooled predictions for respondents in the “No” category, 87 of 165 predictions in the “Not sure” category, and 136 of 220 predictions in the “Yes” category.

The most frequent error for the “No” category was prediction as “Yes” (67 cases). Similarly, 53 “Not sure” predictions and 55 “Yes” predictions were assigned to the “Yes” and “Not sure” categories, respectively. These patterns show that the model had difficulty distinguishing between the three self-reported categories, particularly for respondents who answered “No.”

Table 6
Confusion matrix for Core-feature KNN

Actual category	Predicted: No	Predicted: Not sure	Predicted: Yes
No	22	31	67
Not sure	25	87	53
Yes	29	55	136

Note. Counts represent pooled outer-test predictions across five repetitions. Each respondent contributes five predictions; therefore, counts are repeated predictions rather than unique individuals.

The confusion matrix reinforces the need to interpret aggregate performance alongside class-specific metrics. Although KNN achieved the highest mean macro-F1 among the tested configurations, its errors were unevenly distributed across response categories. The model should therefore be understood as an

exploratory classifier rather than a validated instrument for identifying individual students experiencing academic burnout.

5. Discussion

5.1. Principal Findings

This study explored whether academic and lifestyle survey responses could distinguish three categories of students' self-reported perceptions of academic burnout. Across eight classification approaches and three prespecified feature sets, the highest mean macro-F1 was achieved by KNN using the Core feature set, with a mean of 0.430 and a standard deviation of 0.112 across outer folds. Its mean accuracy was 0.485, compared with 0.436 for the majority-class baseline. Although the best-performing model exceeded the baseline on these measures, the difference was modest, and variability across folds indicates that performance was not consistent across evaluation partitions.

These findings suggest that the available responses contained some information relevant to the three-category outcome, but they do not establish that the model can predict burnout reliably in a new student population. A model can perform better than a simple baseline without achieving the stability or validity required for practical use. In this study, performance was constrained by the small sample, the distribution of outcome categories, the subjective single-item target, and conceptual overlap between some predictors and the outcome.

The results should therefore be understood as a preliminary assessment of predictive feasibility. They demonstrate the application of a consistent comparison procedure to a small survey dataset, rather than the development of a validated prediction instrument. The study does not establish a clinical or psychological assessment method and does not provide sufficient evidence to support individual-level screening.

5.2. Interpreting the Feature-Set Comparison

The comparison between the Full, Core, and Reduced feature sets provides insight into the sensitivity of model performance to predictor selection. The Full feature set included 23 requested predictors from Q2–Q20 and Q22–Q25. The Core feature set retained 19 predictors from Q2–Q20. The Reduced feature set retained 13 predictors from Q2–Q6, Q8–Q10, and Q16–Q20.

The Core feature set produced the highest observed mean macro-F1, at 0.430, compared with 0.410 for the Full feature set and 0.386 for the Reduced feature set. This pattern suggests that including more variables did not necessarily improve classification. Additional predictors may introduce redundant information or noise, and their relationships may be difficult to estimate reliably in a small sample. However, the results do not demonstrate that removing particular questions caused the improvement. These are descriptive comparisons, not statistically established differences.

The lower performance of the Reduced feature set may indicate that the excluded questions contained information associated with respondents' answers to the burnout item. Yet this interpretation requires

caution. Questions about academic and psychological experiences may overlap conceptually with how a student understands or reports burnout. A model that uses related responses may classify the target partly because the predictors and outcome capture overlapping aspects of the same self-reported experience.

Removing some items can help examine the sensitivity of results to predictor composition, but it does not eliminate all conceptual overlap. Consequently, the performance of the Core feature set cannot be interpreted as evidence that the model predicts burnout independently of respondents' contemporaneous perceptions of academic difficulty.

Future studies should define predictor groups before analysis and explain their theoretical rationale. Researchers could distinguish contextual variables, lifestyle variables, and psychological or academic-experience variables. Comparing these groups in a larger dataset would help establish whether classification depends primarily on items that closely resemble the target or whether more independent contextual information contributes useful predictive value.

5.3. Class-Specific Performance and the Meaning of Errors

Overall accuracy alone does not adequately describe the model's behaviour across the three response categories. In the pooled outer-test predictions for the Core KNN model, the F1-score was 0.571 for "Yes," 0.515 for "Not sure," and 0.224 for "No." The corresponding recalls were 0.618, 0.527, and 0.183. Performance differed considerably by category, with the "No" response being particularly difficult to identify.

The pooled confusion matrix illustrates this imbalance. Of the 120 outer-test predictions associated with actual "No" responses across repeated evaluations, 22 were classified as "No," 31 as "Not sure," and 67 as "Yes." In contrast, 136 of the 220 pooled predictions associated with actual "Yes" responses were correctly classified as "Yes." These pooled counts include repeated predictions of the same respondents across five outer cross-validation repeats; they do not represent 505 unique participants.

This pattern has practical significance for interpreting aggregate metrics. The model's moderate macro-F1 reflects its performance across all three categories, but that summary can obscure the particularly weak recognition of respondents who answered "No." A model that frequently confuses "No" with "Yes" would be unsuitable for consequential judgements about students, even if its overall accuracy were higher than the baseline.

The "Not sure" category also deserves attention. It may reflect uncertainty about the meaning of burnout or difficulty evaluating one's own experience, but the current survey does not establish why respondents selected this category. It would therefore be inappropriate to interpret "Not sure" as a measured level of burnout severity or as a clinical intermediate category. Future research could investigate how students understand the term and whether clearer response definitions or validated instruments produce more interpretable outcome categories.

For these reasons, model evaluation in student well-being research should report class-specific precision, recall, and F1 alongside aggregate metrics. It should also consider the consequences of different

classification errors. In an educational setting, incorrectly labelling a student as experiencing burnout could create unnecessary concern or stigma, while failing to recognise a student who needs support could also be harmful. The present study did not evaluate these consequences and provides no basis for selecting an operational decision threshold.

5.4. Small-Sample Validation and Model Uncertainty

The analysis used repeated stratified five-fold outer cross-validation, with five repeats, and hyperparameter selection within the training portion of each outer fold. This design separates hyperparameter tuning from the corresponding outer-fold evaluation and reduces the risk of evaluating a model on observations used directly to select its hyperparameters. Fitting imputation, scaling, and categorical encoding within training partitions also helps prevent information from the outer test partition from influencing preprocessing.

Nevertheless, careful validation design cannot fully compensate for limited data. With 101 responses divided among three outcome categories, each test fold contains relatively few observations from at least some categories. The inclusion or exclusion of a small number of respondents can therefore materially affect fold-level performance. The observed standard deviations across outer folds illustrate variation under the evaluation procedure, although they should not be interpreted as confidence intervals or independent estimates of population uncertainty.

Repeated cross-validation uses the available sample efficiently, but repeated partitions do not create new participants. The same respondents contribute to multiple test folds across repeats, so pooled predictions and fold-level summaries are dependent. The reported results are descriptive evidence about model behaviour under this resampling procedure. They do not demonstrate that performance will be reproduced in another institution, cohort, or period.

The comparison of eight approaches and three feature sets also creates a risk of overemphasising the highest observed score. The Core KNN model is the best configuration among the reported comparisons, but its ranking should not be treated as proof that this algorithm is generally superior for predicting student burnout. Small-sample model rankings can be unstable, and the best observed configuration may reflect characteristics specific to the available sample. Independent external validation is necessary to determine whether a selected model performs similarly on new data.

A stronger follow-up study would recruit a larger and more diverse sample, predefine the primary evaluation metric, document the full analysis pipeline, and evaluate the selected model on an independent dataset. If practical use were eventually considered, further work would also need to examine calibration, error costs, subgroup performance, and whether predictions offer meaningful value beyond simpler and more transparent approaches.

5.5. Implications for Student Support and Educational Research

The findings do not justify using machine learning to screen students or determine who is experiencing burnout. Their more appropriate implication is methodological: researchers can use exploratory

classification to investigate information contained in student surveys, provided that they report uncertainty and avoid overstating predictive capability. In this study, the modest improvement over the majority-class baseline and the weak recognition of the “No” category reinforce the need to distinguish exploratory modelling from a validated support system.

Student well-being is multidimensional, and a single response may not capture changes in exhaustion, engagement, motivation, or personal circumstances over time. Institutions seeking to understand student experiences should prioritise validated instruments, voluntary participation, appropriate support pathways, and careful interpretation of self-report data. Any use of predictive analytics in this area should be subject to appropriate ethical review, privacy safeguards, and evaluation of potential harms. Students should not be labelled or denied support on the basis of an unvalidated model.

The present dataset also has limited relevance to claims about educational technology. Although predictive analytics may be used in educational settings, the survey did not directly measure learning-management-system activity, digital course participation, technology-supported learning design, or students’ experiences with educational platforms. The study therefore cannot establish whether educational-technology use is associated with the outcome or whether technology-based interventions could improve student well-being.

A future educational-technology study could examine whether students’ experiences of online course design, digital workload, platform accessibility, or technology-mediated communication relate to validated measures of well-being. Such a study would need to measure the relevant digital-learning context explicitly rather than assume that general screen time represents educational-technology use.

5.6. Limitations and Future Research

Several limitations should guide interpretation. First, the dataset contained only 101 responses, limiting the precision and generalisability of the results. Second, the target was a single self-reported item rather than a validated multidimensional measure. Third, the sample was collected through a survey link shared in student groups and social media, so it should not be assumed to represent the broader university student population. Fourth, some predictors may overlap conceptually with the target, making it difficult to distinguish independent contextual information from related self-reported experiences.

Fifth, the evaluation used internal cross-validation rather than an independent external test dataset. Although the nested procedure helps separate hyperparameter tuning from performance evaluation, it cannot establish generalisability to other populations. Sixth, the analysis compared multiple models and feature sets, so the highest observed score should be treated as a result within this comparison rather than definitive evidence that one algorithm is best. Finally, the survey did not directly measure educational-technology use and cannot support conclusions about the effects of digital learning environments.

Future research should use validated burnout measures, larger and more diverse samples, and clearly prespecified research questions. It should distinguish predictors available before the outcome from variables that may measure similar contemporaneous experiences. An independent validation sample should be used to assess generalisation. Researchers should also report class-specific errors, uncertainty,

and potential consequences of misclassification. Any proposed use in student support would require a separate assessment of validity, fairness, privacy, acceptability, and potential harm.

6. Conclusion and Recommendations

This exploratory study examined whether academic and lifestyle survey responses could distinguish students' self-reported perceptions of current academic burnout. Using 101 responses, eight classification approaches, and three feature sets, the analysis found that KNN with the Core feature set achieved the highest mean macro-F1 score (0.430, SD = 0.112) and mean accuracy of 0.485 (SD = 0.116). These results represented a modest improvement over the majority-class baseline accuracy of 0.436, but performance varied across evaluation folds and outcome categories.

The class-specific results showed that the selected model identified "Yes" and "Not sure" responses more effectively than "No" responses. This uneven performance, together with the small sample and the use of a single self-reported target item, limits the practical interpretation of the findings. The model should not be treated as a validated burnout assessment or used to make decisions about individual students.

Future research should recruit a larger and more diverse sample, use validated measures of academic burnout, and evaluate models on independent data. Researchers should also examine subgroup performance, calibration, and the consequences of different classification errors. Any future study involving student data should address applicable research ethics requirements before collecting responses and should protect participants' privacy.

In conclusion, machine-learning classification may offer a way to explore patterns in student survey responses, but the present results provide only preliminary evidence of predictive feasibility. Further methodological validation and a clearer connection to educational practice are necessary before such approaches can support student well-being initiatives. Because this study did not directly measure educational-technology use, it cannot establish implications for technology-supported learning or demonstrate that educational technology can identify or reduce academic burnout.

References

1. Auerbach RP, Alonso J, Axinn WG, Cuijpers P, Ebert DD, Green JG, Hwang I, Kessler RC, Liu H, Mortier P, Nock MK, Pinder-Amaker S, Sampson NA, Aguilar-Gaxiola S, Al-Hamzawi A, Andrade LH, Benjet C, Caldas-de-Almeida JM, Demyttenaere K, Bruffaerts R (2016) Mental disorders among college students in the World Health Organization World Mental Health Surveys. Psychol Med 46(14):2955–2970. https://doi.org/10.1017/S0033291716001665
2. Bedewy D, Gabriel A (2015) Examining perceptions of academic stress and its sources among university students: The Perception of Academic Stress Scale. Health Psychol Open 2(2). Article 2055102915596714. https://doi.org/10.1177/2055102915596714
3. Boni RADS, Paiva CE, de Oliveira MA, Lucchetti G, Fregnani JH, T. G, Paiva BSR (2018) Burnout among medical students during the first years of undergraduate school: Prevalence and associated factors.

PLoS ONE 13(3) Article e0191746. https://doi.org/10.1371/journal.pone.0191746

4. Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
5. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018
6. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F (2018) Learning from imbalanced data sets. Springer. https://doi.org/10.1007/978-3-319-98074-4
7. Hara Y, Furuta S, Sukegawa A, Fujimaki K, Hisano S (2026) Systematic review of student burnout instruments: No measures currently meet COSMIN standards for recommendation. BMC Psychol 14:964. https://doi.org/10.1186/s40359-026-04570-x
8. Iyortsuun NK, Kim S-H, Jhon M, Yang H-J, Pant S (2023) A review of machine learning and deep learning approaches on mental health diagnosis. Healthcare 11(3). Article 285. https://doi.org/10.3390/healthcare11030285
9. Jagodics B, Szabó É (2023) Student burnout in higher education: A demand-resource model approach. Trends Psychol 31:757–776. https://doi.org/10.1007/s43076-021-00137-4
10. Kaggwa MM, Kajjimu J, Sserunkuma J, Najjuka SM, Atim LM, Olum R, Tagg A, Bongomin F (2021) Prevalence of burnout among university students in low- and middle-income countries: A systematic review and meta-analysis. PLoS ONE 16(8) Article e0256402. https://doi.org/10.1371/journal.pone.0256402
11. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
12. Maslach C, Schaufeli WB, Leiter MP (2001) Job burnout. Ann Rev Psychol 52:397–422. https://doi.org/10.1146/annurev.psych.52.1.397
13. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html
14. Rosales-Ricardo Y, Rizzo-Chunga F, Mocha-Bonilla J, Ferreira JP (2021) Prevalence of burnout syndrome in university students: A systematic review. Salud Mental 44(2):91–102. https://doi.org/10.17711/SM.0185-3325.2021.013
15. Rosenman R, Tennekoon V, Hill LG (2011) Measuring bias in self-reported data. Int J Behav Healthc Res 2(4):320–332. https://doi.org/10.1504/IJBHR.2011.043414
16. Schaufeli WB, Martinez IM, Pinto AM, Salanova M, Bakker AB (2002) Burnout and engagement in university students: A cross-national study. J Cross-Cult Psychol 33(5):464–481. https://doi.org/10.1177/0022022102033005003
17. Shatte ABR, Hutchinson DM, Teague SJ (2019) Machine learning in mental health: A scoping review of methods and applications. Psychol Med 49(9):1426–1448. https://doi.org/10.1017/S0033291719000151

18. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manag 45(4):427–437. https://doi.org/10.1016/j.ipm.2009.03.002
19. Vabalas A, Gowen E, Poliakoff E, Casson AJ (2019) Machine learning algorithm validation with a limited sample size. PLoS ONE 14(11) Article e0224365. https://doi.org/10.1371/journal.pone.0224365
20. Varoquaux G (2018) Cross-validation failure: Small sample sizes lead to large error bars. NeuroImage 180:68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Figures

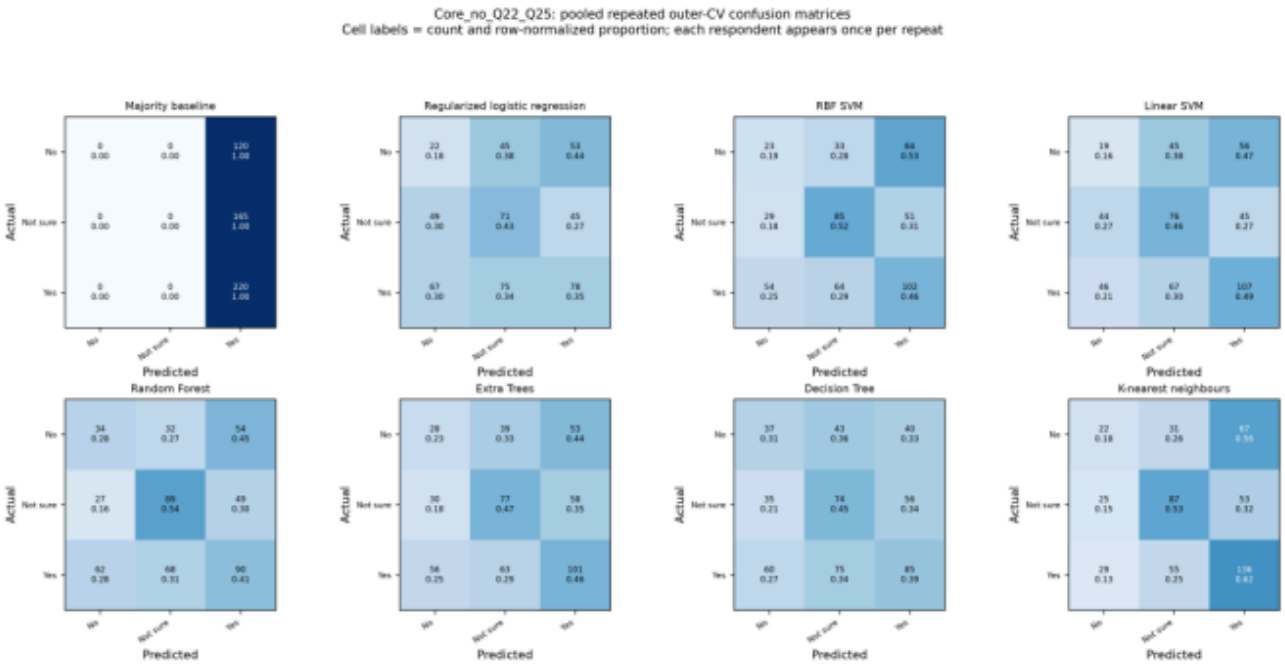


Figure 1

Confusion matrix for the k-nearest neighbours classifier using the core feature set. The matrix presents the pooled classification results across repeated stratified five-fold cross-validation, showing the actual and predicted categories of self-reported academic burnout (No, Not sure, and Yes).

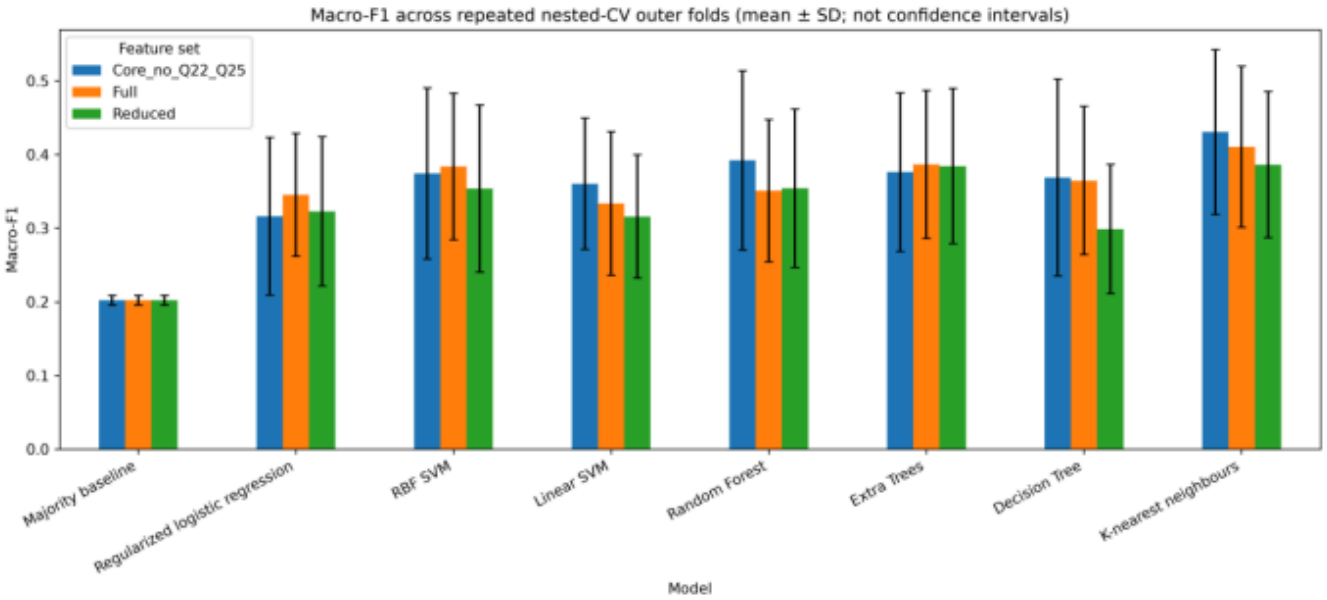


Figure 2

Comparison of classification performance across the three feature sets. The figure compares the performance of the evaluated machine-learning classifiers using macro-averaged F1 scores, highlighting differences in classification performance across feature configurations.



Figure 3

Overview of the machine-learning research workflow. The figure illustrates the main stages of the study, including survey data preparation, feature-set construction, preprocessing, model training and hyperparameter selection, repeated stratified five-fold cross-validation, and performance evaluation.