RESEARCH ARTICLE

How Do Multiple-Choice Assessments Reflect Student Performance? An Educational Data Mining Analysis Using Machine Learning Classifiers

Niusha Shafiabady[1,2,*], **Darius Pfitzner**[3], **Talha Nawaz**[4], **Seb Dianati**[5] and **Mark Brady**[6]

[1] Women in AI for Social Good Lab, Discipline of Information Technology, Australian Catholic University, North Sydney, NSW 2060, Australia
[2] Faculty of Science and Technology, Charles Darwin University, Sydney, NSW 2000, Australia
[3] Australian Institute of Business, Adelaide, SA 5000, Australia
[4] Department of Electrical Engineering-Electronics-Information Technology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen 91054, Germany
[5] Learning and Teaching Innovation, University of Southern Queensland, Springfield, QLD 4300, Australia
[6] Law, Faculty of Arts and Society, Charles Darwin University, Darwin, NT 0909, Australia

Abstract

Multiple-choice questions (MCQs) are widely used in higher education because they efficiently assess knowledge and a broad range of cognitive skills across disciplines. This study examines how MCQ-based assessments contribute to predicting final student grades, using data from 2,390 students (comprising 26,698 non-missing attribute values) enrolled in sixteen offerings of three undergraduate and postgraduate economics units at Charles Darwin University (CDU) between 2016 and 2019. Six machine learning classifiers were trained on assessment marks, assessment weightings, final examination performance, and the proportion of MCQs in the examination. Feature importance analysis was then used to quantify the contribution of each assessment component. The study compares MCQ and task-based examination questions, online MCQ and task-based assessments, and online and written MCQ components. Support Vector Machines achieved the highest accuracy (96.07%), followed by XGBoost (93.76%). Across the higher-performing classifiers, final examination performance and the task-based assessment were the strongest predictors of final grades. Online and written MCQ components contributed less strongly, although MCQs remain useful indicators of foundational knowledge. These findings support balanced assessment designs that combine MCQs with task-based activities, and they demonstrate how explainable AI can inform assessment analytics within Educational Data Mining.

Keywords: Educational data mining, learning analytics, multiple-choice questions, student performance prediction, assessment analytics, artificial intelligence, machine learning, explainable AI.

Submitted: 30 June 2026
Revised: 11 August 2026
Accepted: 26 August 2026
Published: 23 October 2026

Vol. 2, **No.** 2, 2026.
10.62762/TEDM.2026.107680

***Corresponding author:**
✉ Niusha Shafiabady
niusha.shafiabady@acu.edu.au

Citation
Shafiabady, N., Pfitzner, D., Nawaz, T., Dianati, S., & Brady, M. (2026). How Do Multiple-Choice Assessments Reflect Student Performance? An Educational Data Mining Analysis Using Machine Learning Classifiers. *ICCK Transactions on Educational Data Mining*, 2(2), 85–101.

© 2026 ICCK (Institute of Central Computation and Knowledge)

1 Introduction

Educational Data Mining (EDM) and Learning Analytics increasingly employ Artificial Intelligence (AI) techniques to analyse student learning behaviours, assessment outcomes, and academic performance. Educational Data Mining aims to extract meaningful patterns from educational data to better understand learning processes, predict academic outcomes, and support evidence-based educational decision-making. Within this context, the present study investigates assessment dynamics by applying multiple AI classifiers to predict student grades based on assessment scores, final examination results, and the proportional contribution of MCQs to overall grades. By examining feature importance across different classifiers, the study contributes to the growing body of Educational Data Mining research concerned with understanding how specific assessment practices influence student success. More broadly, this aligns with recent calls for EDM systems that extend beyond predictive accuracy to emphasise interpretability, transparency, and educational relevance, enabling AI models to better support evidence-informed educational decision-making [1].

A crucial aspect of this research is examining the importance of each input, particularly in understanding the nuanced impact that MCQs have on overall results. MCQs are widely used in both formative and summative assessment and provide an efficient means of evaluating student learning [2, 3] in closed- and open-book environments [4]. In higher education, discussion around MCQs continues to focus on their appropriate application and validity [5–9]. Building on the widespread use of MCQs and their recognised value in assessing disciplinary knowledge, this study applies several AI classifiers to student assessment data from a regional Australian university to investigate how different assessment components contribute to final academic outcomes.

The main objective of this study is to address questions regarding the influence of MCQ assessments compared to task-based examination questions. Furthermore, it aims to investigate the specific effects of online MCQ assessments on overall grades. The study also examines whether online MCQ assessments and written examination MCQ components differ in their predictive influence on student performance. This investigation provides important insights into how MCQ assessments relate to students' academic outcomes. The primary research questions (RQs) are:

RQ1: To what extent do MCQ components influence examination performance compared with task-based examination questions?

RQ2: How does the contribution of online MCQ assessments compare with task-based assessments in predicting final student grades?

RQ3: Does the predictive influence of online MCQ assessments differ from that of written examination MCQ components?

This study makes three contributions to the Educational Data Mining literature. First, it investigates the relative predictive influence of MCQ and task-based assessments using a large real-world dataset comprising records from 2,390 students across sixteen unit offerings. Second, it compares multiple AI classification techniques to evaluate their effectiveness in predicting student performance. Third, feature importance analysis provides explainable insights into the relative contribution of different assessment components to final academic outcomes, supporting evidence-based assessment design and learning analytics practice.

2 Literature Review and Research Context

2.1 Multiple-Choice Questions in Higher Education Assessment

Multiple-choice questions remain one of the most widely used assessment formats in higher education because of their efficiency, scalability, and ability to provide broad curriculum coverage [3]. Assessment outcomes from MCQs and other assessment instruments are commonly aggregated to determine final grades, such as High Distinction (HD), Pass (P), and Fail (F), which serve as indicators of overall student performance within a unit of study [8]. MCQs are particularly effective for assessing the recall and recognition of factual knowledge and therefore remain common in disciplines where foundational understanding is essential [5].

Beyond their role as summative assessment tools, MCQs are increasingly used to support learning. Research has shown that incorporating MCQs into classroom activities can improve engagement and learning outcomes [2]. Similarly, written justifications accompanying MCQ responses can encourage deeper reasoning and conceptual understanding [10]. Improvements in summative assessment performance have also been reported when MCQs are used as

formative learning activities during lectures [11]. Collectively, these findings suggest that MCQs can serve both assessment and learning functions.

The educational value of MCQs, however, has been the subject of longstanding debate [7]. While some researchers argue that MCQs provide an efficient and reliable means of assessing student knowledge, others suggest that they may inadequately capture deeper forms of understanding and higher-order thinking [12, 13]. The relationship between multiple-choice and constructed-response assessments has also produced mixed findings. Some studies report weak relationships between the two formats, suggesting that they assess different dimensions of knowledge and performance [14], whereas others conclude that the two formats largely measure the same underlying construct [15].

Empirical evidence from economics and science education further illustrates this complexity. Kennedy and Walstad examined, from an economist's perspective, how multiple-choice and constructed-response scores should be combined in economics testing [16]. Becker and Johnston [17] found no significant relationship between performance on multiple-choice and essay-based economics assessments, concluding that the two formats evaluate distinct aspects of student learning. Similar conclusions were reported by Dufresne et al. [18], who questioned whether one assessment format could reliably predict performance in another. A meta-analysis of 67 empirical studies by Rodriguez found considerable variation in the correlations reported across the literature and identified stem equivalence between MCQ and constructed-response items as a key factor influencing the observed relationships [19]. These findings suggest that the educational value of MCQs depends not only on their format but also on their design and implementation.

2.2 MCQs in Digital and Online Learning Environments

The widespread adoption of online learning environments has generated renewed interest in the role of MCQ-based assessments within digital education. Online assessment systems offer flexibility, accessibility, and learner autonomy, potentially increasing student confidence and engagement [4, 20]. Evidence from introductory economics and accounting courses further demonstrates that the active use of online MCQ resources is positively associated with examination performance [21]. These advantages have contributed to the growing use of MCQs in online and blended learning environments.

At the same time, assessment performance is influenced by factors beyond assessment format alone. Test anxiety and perceived stress have been shown to affect student performance in online assessment contexts [22]. Similarly, the self-regulated learning strategies that students adopt in online environments—including time management, metacognitive monitoring, and help-seeking behaviours—are systematically associated with academic achievement [23]. Although these factors are not directly examined in the present study, they represent important contextual variables that should be considered when interpreting assessment outcomes.

Despite their advantages, MCQs in online environments are not without limitations. Some studies report lower performance and higher withdrawal rates in online courses [24], suggesting that online MCQs may need to be complemented with practical or task-based assessments. Other researchers have highlighted issues such as misleading distractors, overconfidence, and inflated perceptions of learning [25–27]. Maintaining student motivation and engagement in online learning environments therefore remains a significant challenge [28]. Research further suggests that differences in self-regulated learning strategies between online and blended learners are associated with differences in academic performance, underscoring the importance of assessment designs that support varied learning approaches [29].

To address these challenges, online learning environments increasingly incorporate feedback, guidance, and interactive learning supports that help sustain engagement and improve learning outcomes [30]. Effective assessment design also requires balancing assessment difficulty with meaningful measurement of learning [20]. Researchers have further argued that poorly designed MCQ items may distort student achievement measures, emphasising the need for rigorous item analysis and quality assurance procedures [8]. These considerations are particularly relevant within Educational Data Mining research, where assessment data are frequently used as indicators of learning outcomes and predictors of future academic performance.

Overall, the literature suggests that assessment performance is influenced not only by assessment type

but also by the broader learning environment in which assessment occurs.

2.3 Assessment Design, Higher-Order Learning, and Educational Data Mining

Recent research has increasingly focused on whether MCQs can be designed to support higher-order learning rather than merely assess factual recall. Butler argues that well-designed MCQs can simultaneously support learning and assessment objectives when questions are appropriately targeted and aligned with intended learning outcomes [6]. Similarly, Nicol [31] and Wilkie et al. [32] highlight the potential of MCQs to encourage learner self-regulation and differentiate student performance across levels of cognitive complexity.

Despite these advantages, student perceptions do not always align with educational intentions. Scouller found that students often associate MCQs with lower-order cognitive processing and consequently adopt less sophisticated study strategies when preparing for such assessments [33]. At the same time, comparisons between MCQs and Modified Essay Questions suggest that carefully designed MCQs can provide reliable measures of factual knowledge while reducing grading burden and maintaining consistency [9]. These findings highlight the importance of assessment design in shaping both learning behaviour and assessment outcomes.

Consequently, understanding how different assessment components contribute to student outcomes remains an important challenge for both assessment researchers and learning analytics practitioners. While the educational literature has extensively debated the strengths and limitations of MCQ assessments, considerably less attention has been devoted to understanding their predictive contribution to academic performance using Educational Data Mining techniques. As higher education institutions increasingly collect large-scale assessment data, opportunities emerge to apply machine learning and learning analytics approaches to evaluate the relative influence of different assessment components on student success. The present research contributes to this emerging area by combining assessment analytics with AI-based classification methods to examine how MCQ and task-based assessments relate to final academic outcomes.

2.4 Educational Data Mining and Assessment Analytics

Educational Data Mining and Learning Analytics provide useful approaches for examining how students engage with learning activities, assessment tasks, and academic progression. Recent advances in Machine Learning (ML) and AI have expanded the capacity of educators and researchers to analyse large-scale educational datasets, identify patterns in student performance, and support evidence-based educational decision-making. Jordan and Mitchell [34] highlight the growing role of machine learning in extracting meaningful insights from complex data, while Chen and Guestrin [35] demonstrate how scalable tree-boosting algorithms can efficiently identify predictive patterns in large, high-dimensional datasets. Systematic reviews of supervised data mining techniques applied to student exam performance prediction further confirm the broad applicability of such classification approaches in educational contexts [36]. These methods can support the analysis of assessment outcomes and provide insights into the factors associated with student success.

Within assessment analytics, AI-based models can be used not only to predict student performance but also to evaluate the relative contribution of different assessment components. This is particularly relevant when assessment designs combine MCQs, task-based assessments, and final examinations. Hu [37] emphasises the role of AI in evaluating teaching effectiveness, which aligns with the aim of using predictive modelling to understand how assessment structures relate to academic outcomes. Similarly, research on multiple-choice testing suggests that data-driven analysis can help clarify how MCQs contribute to both learning and assessment outcomes [30].

Many approaches may be used to evaluate the utility of MCQ-style assessments, ranging from descriptive statistics to more advanced machine learning techniques. Machine learning is an interdisciplinary field that draws on statistics, computer science, optimisation, and cognitive science to model patterns in data and support prediction [38, 39]. Systematic reviews confirm that machine learning models—including ensemble and classification-based approaches—have made substantial contributions to student academic performance prediction across diverse educational datasets [40]. In this study, classification methods are particularly relevant because they allow student outcomes to be grouped

into defined categories based on assessment-related input features. Supervised classification algorithms have been widely used to assign observations to predefined classes, making them suitable for student performance prediction and grade classification tasks [41, 42].

In the context of this study, assessment data are treated not simply as final marks, but as evidence of how different assessment components contribute to student performance. By comparing model accuracy and feature importance across multiple AI classifiers, the study examines whether MCQ-based assessments, task-based assessments, and written examination components differ in their predictive influence on final grades. This approach supports the broader aims of EDM by using educational data to inform assessment design, learning analytics practice, and evidence-based curriculum improvement.

3 Machine Learning Models for Student Performance Prediction

Recent advances in ML and AI have created new opportunities for analysing educational data and evaluating assessment practices. Educational Data Mining and Learning Analytics increasingly employ AI techniques to identify patterns in student performance, predict academic outcomes, and support evidence-based decision-making in higher education [34, 43]. These developments have enabled researchers to move beyond traditional statistical analyses and examine complex relationships within large educational datasets. In the present study, machine learning techniques are employed to examine the relationship between MCQ assessments, task-based assessments, examination performance, and final student grades. By comparing predictive accuracy and feature importance across multiple classifiers, the study seeks to better understand how different assessment components contribute to academic success.

3.1 Selection of Classification Models

To investigate the relationship between assessment components and student outcomes, this study employs six machine learning classifiers representing different modelling approaches: Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Decision Trees (DT), Extremely Randomized Trees (Extra Trees), Extreme Gradient Boosting (XGBoost), and Logistic Regression. These models represent margin-based, instance-based, tree-based, ensemble, boosting-based, and statistical classification approaches, respectively. Comparing predictive performance and feature importance across these methods provides a basis for examining whether patterns in the relative predictive contribution of assessment components are consistent across different modelling approaches.

Machine learning classification has become one of the most widely applied approaches for educational prediction tasks because it enables observations to be assigned to predefined categories based on historical patterns within the data [41, 42]. The classifiers selected for this study have demonstrated strong performance across a variety of prediction domains and collectively provide a balanced comparison of different learning strategies. More broadly, machine learning methods have been widely applied to student performance prediction and related tasks in educational data mining and learning analytics, illustrating the flexibility of AI techniques for modelling multidimensional educational datasets [43].

3.2 Support Vector Machines

Support Vector Machines (SVMs) are among the most widely used supervised machine learning algorithms for classification and prediction tasks. Originally developed by Cortes and Vapnik [44], SVMs construct an optimal separating hyperplane that maximises the margin between classes, thereby improving generalisation performance on unseen data [45]. Their ability to operate effectively in high-dimensional spaces has contributed to their widespread adoption in data mining, pattern recognition, and educational analytics applications [46].

Previous studies have demonstrated the effectiveness of SVMs across a wide range of classification problems, often outperforming alternative methods when dealing with complex feature interactions [42, 47]. For a binary kernel SVM, the decision function is expressed as

$$f(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i \, y_i \, K(\mathbf{x}_i, \mathbf{x}) + b\right), \qquad (1)$$

where $\mathbf{x}$ is the input feature vector to be classified, $\mathbf{x}_i$ $(i = 1, \ldots, n)$ are the n training samples (only the support vectors have $\alpha_i > 0$), α_i are the Lagrange multipliers (support-vector coefficients), $y_i \in \{-1, +1\}$ are the class labels, $K(\cdot, \cdot)$ is the kernel function (for a linear SVM, $K(\mathbf{x}_i, \mathbf{x}) = \mathbf{x}_i^\top \mathbf{x}$), and b is the bias term. For multi-class grade prediction,

binary SVMs are combined through a one-versus-one or one-versus-rest decomposition. By maximising the margin between classes, SVMs aim to improve classification accuracy and reduce overfitting [44].

3.3 K-Nearest Neighbours

K-Nearest Neighbours (KNN) is an instance-based learning algorithm that classifies observations according to the characteristics of neighbouring examples within the feature space. Unlike many machine learning methods, KNN does not construct an explicit predictive model during training. Instead, classification decisions are based on the similarity between new observations and previously observed instances.

The principal tuning parameter in KNN is the number of neighbours, denoted by k. Smaller values of k may produce highly flexible decision boundaries, whereas larger values may oversimplify class distinctions and reduce predictive sensitivity [48]. Despite its conceptual simplicity, KNN has demonstrated effectiveness across numerous classification applications and remains attractive because of its adaptability to different data structures [49].

3.4 Decision Trees and Ensemble Tree Methods

3.4.1 Decision Trees

Decision Trees (DTs) are among the most interpretable machine learning techniques. They recursively partition data into increasingly homogeneous subsets using a sequence of decision rules based on predictor variables. The resulting tree structure provides a transparent representation of the classification process, making DTs particularly attractive in educational research where model interpretability is important [50].

Each internal node represents a feature-based decision, branches represent alternative outcomes, and terminal leaf nodes represent predicted class labels. Although Decision Trees are easy to interpret, their predictive performance may decline when data relationships become highly complex, and they can be prone to overfitting if not properly controlled.

3.4.2 Extremely Randomized Trees

Extremely Randomized Trees (Extra Trees) is an ensemble learning method that constructs multiple decision trees and combines their predictions to improve generalisation performance. Compared with conventional decision-tree learning, Extra Trees introduces additional randomisation during tree construction by randomly selecting candidate features and split thresholds. The predictions of the individual trees are then aggregated to produce the final classification result. This randomisation can reduce variance and limit overfitting while retaining the ability of tree-based models to capture non-linear relationships and interactions among predictor variables [51].

3.4.3 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a boosting-based ensemble technique that combines multiple weak learners, typically shallow decision trees, to create a stronger predictive model. Unlike Random Forest, which constructs trees independently, boosting builds trees sequentially, with each new tree fitted to reduce the residual errors (gradients of the loss) of the current ensemble. XGBoost further incorporates regularisation of tree complexity, shrinkage, and column subsampling, which help control overfitting and enable efficient training on large datasets [35]. Through the aggregation of multiple weak learners, XGBoost often achieves higher predictive accuracy than individual classifiers and can capture non-linear relationships and interactions among predictor variables.

3.5 Logistic Regression

Logistic Regression is a statistical classification model that estimates the probability that an observation belongs to each outcome category as a function of the predictor variables [52]. For a grade outcome y with C categories, the multinomial logistic regression model is

$$P(y = c \mid \mathbf{x}) = \frac{\exp\left(\beta_{c0} + \boldsymbol{\beta}_c^\top \mathbf{x}\right)}{\sum_{j=1}^{C} \exp\left(\beta_{j0} + \boldsymbol{\beta}_j^\top \mathbf{x}\right)}, \quad c = 1, \ldots, C, \tag{2}$$

where $\mathbf{x}$ is the vector of predictor variables, β_{c0} is the intercept and $\boldsymbol{\beta}_c$ is the coefficient vector for category c, and the predicted grade is the category with the highest probability [53].

Because the log-odds are modelled as a linear function of the predictors, Logistic Regression provides a transparent statistical baseline against which the non-linear classifiers can be compared. In the context of educational assessment analytics, it helps identify the strength and direction of associations between assessment components and overall academic performance.

3.6 Summary of Selected Models

The machine learning models selected for this study represent a range of classification approaches commonly used in predictive analytics. Support Vector Machines provide a margin-based learning framework for identifying class boundaries, while K-Nearest Neighbours uses an instance-based approach in which observations are classified according to neighbouring cases in the feature space. Decision Trees provide an interpretable rule-based structure, whereas Extremely Randomized Trees use an ensemble of randomised decision trees to improve generalisation. XGBoost adopts a sequential boosting strategy in which successive trees are fitted to improve the current ensemble. Logistic Regression is included as a statistical baseline for modelling the relationship between assessment variables and grade outcomes.

The use of multiple modelling approaches serves two important purposes. First, it reduces the likelihood that findings are dependent on the assumptions or limitations of a single classifier. Second, it enables a comparison of feature importance across different learning paradigms, providing a more comprehensive understanding of how assessment components contribute to student outcomes. Consistent patterns observed across multiple classifiers provide stronger evidence regarding the relative influence of MCQ assessments, task-based assessments, and examination performance.

From an Educational Data Mining perspective, the objective is not only to identify the classifier with the highest predictive accuracy, but also to understand the educational significance of the variables driving model performance. By examining both predictive capability and feature importance, this study seeks to determine the extent to which different assessment components contribute to final grades and whether MCQ assessments provide comparable predictive value to task-based assessments. Balancing predictive performance with explainability reflects an emerging direction within Educational Data Mining that encourages transparent and trustworthy AI systems capable of supporting human-centred educational decision-making [1].

4 Dataset and Methodology

4.1 Dataset Description

The dataset used in this study was compiled from undergraduate and postgraduate economics units offered at Charles Darwin University (CDU) between 2016 and 2019. To obtain individual student performance on the multiple-choice question (MCQ) component of examinations, assessment records were manually extracted from archived paper-based examinations and institutional grade reports. Only students who completed a unit and received a final semester grade were included in the analysis.

Data were collected from sixteen unit offerings (three economics units delivered across multiple semesters) during the four-year study period. In total, 2,390 student records were obtained. Following the inclusion of both original assessment variables and derived variables, the final dataset comprised 26,698 non-missing attribute values across all records. Table 1 summarises the assessment structure and composition of the dataset.

Within the analysed units, Assessments 1 and 2 consisted of online MCQ-based quizzes administered during the semester. Assessment 3 primarily comprised task-based assessment activities, while the final examination was conducted in a face-to-face setting and included both MCQ and task-based question components. In this study, a *unit* refers to a subject taught over a twelve-week teaching period at CDU. This assessment structure provided a suitable environment for investigating the relative contribution of different assessment formats to final student outcomes.

4.2 Assessment Structure and Variables

A range of assessment-related variables was used as input features for the machine learning models. These variables captured assessment performance, assessment weighting, unit characteristics, and examination structure. The complete set of input variables used in the analysis is presented in Table 2.

The variables included marks obtained in the two online MCQ assessments, Assessment 1 and Assessment 2, marks obtained in the task-based Assessment 3, final examination performance, the weighting assigned to each assessment component, and the relative proportion of MCQ questions within the final examination. Additional contextual variables included the year of delivery, semester, and unit code.

The target variable for all classification models was the final grade category awarded to each student. The objective was to determine the extent to which individual assessment components contributed to predicting final academic outcomes and to examine the relative influence of MCQ-based and task-based

Table 1. Summary of assessment structures across the analysed units.

		Final exam details			Mid-semester test details	
Unit	**Year**	**No. of exam marks**	**Exam weight (%)**	**MCQ portion of exam (%)**	**No. of tests**	**Total weight (%)**
Principles of Economics	2016	327	60	40	3	20
	2017	292	60	57	3	20
	2018	190	50	57	3	20
	2019	116	50	20	3	20
Economics for Managers	2016	242	50	20	3	20
	2017	185	50	20	3	20
	2018	60	50	20	3	20
	2018	74	50	40	3	10
Business Economics	2016	154	50	40	3	10
	2016	176	50	20	3	10
	2017	132	50	40	3	10
	2017	78	50	30	3	10
	2018	78	50	20	3	10
	2018	66	50	20	3	10
	2019	132	50	20	3	10
	2019	88	50	20	3	10
Total		**2,390**				

Table 2. Input variables used for grade classification.

Input feature	**Description**
Year	Year of the examination
Unit Code	Code of the unit (subject)
Semester	Semester in which the unit was taught
Assessment 1	Marks in online MCQ-based mid-semester Assessment 1
Assessment 1 Weight (%)	Weight of Assessment 1 towards the final grade
Assessment 2	Marks in online MCQ-based mid-semester Assessment 2
Assessment 2 Weight (%)	Weight of Assessment 2 towards the final grade
Assessment 3	Marks in the task-based Assessment 3
Assessment 3 Weight (%)	Weight of Assessment 3 towards the final grade
Final Exam	Written final examination containing both task-based and MCQ questions
Exam Weight (%)	Weight of the final examination towards the final grade
MCQ Portion of Exam (%)	Relative percentage of MCQs in the written examination

assessments on student performance.

4.3 Experimental Design and Model Development

Six machine learning algorithms were employed to classify student grade outcomes: Decision Tree (DT), Extremely Randomized Trees (Extra Trees), K-Nearest Neighbours (KNN), Logistic Regression, Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost). These models were selected to represent several complementary classification paradigms, including margin-based learning, instance-based learning, tree-based learning, randomised tree ensembles, gradient boosting, and statistical classification. Evaluating multiple modelling approaches provides a more robust basis for assessing the relative contribution of assessment components to final grade outcomes.

All models were implemented in Python using standard machine learning libraries and trained using the assessment variables described in Table 2. The dataset was partitioned into training and testing subsets, allowing model performance to be evaluated on previously unseen data. Following model development, feature importance analysis was conducted to identify the variables contributing most strongly to grade prediction. Feature importance values were extracted using the importance measures available for each classifier and were normalised within each model so that the most important feature received a score of 1. This normalisation facilitates comparison of the relative importance and ranking of assessment variables within each classifier. Because feature importance may be derived differently across modelling approaches, the absolute

Table 3. Classification accuracy and normalised feature-importance scores for the six evaluated classifiers.

Feature	SVM	XGBoost	KNN	Extra Trees	DT	Logistic Regression
Accuracy	**96.07%**	**93.76%**	**88.45%**	**87.76%**	**82.22%**	**69.75%**
Feature importance score						
Year	0.006040	0.165060	0.000000	0.090000	0.082280	0.079040
Unit Code	0.000000	0.050930	0.000000	0.067540	0.008730	0.322400
Semester	0.002890	0.128050	0.007320	0.046750	0.041350	0.728090
Assessment 1	0.134520	0.253120	0.090040	0.189430	0.137990	0.000000
Assessment 1 Weight (%)	0.002360	0.000000	0.003660	0.000000	0.000000	0.213950
Assessment 2	0.154230	0.200410	0.089200	0.170080	0.112430	0.044690
Assessment 2 Weight (%)	0.002360	0.000000	0.003660	0.000000	0.000000	0.213950
Assessment 3	0.715190	0.521420	0.660380	0.327630	0.352420	0.018260
Assessment 3 Weight (%)	0.002360	0.799870	0.003660	0.083790	0.055410	0.171390
Final Exam	1.000000	1.000000	1.000000	1.000000	1.000000	0.002720
Exam Weight (%)	0.002360	0.000000	0.003660	0.086300	0.055960	1.000000
MCQ Portion of Exam (%)	0.021020	0.137360	0.106080	0.216070	0.052740	0.116750

normalised scores should not be interpreted as directly comparable between classifiers. Examining the feature-importance patterns across multiple classifiers also improves model interpretability by identifying assessment variables that consistently emerge as influential predictors across different modelling approaches. Such explainable methods have become increasingly important for ensuring that AI models remain transparent, interpretable, and practically useful within educational settings [1]. Related work in educational data mining has similarly combined machine-learning prediction with feature-importance and explainability techniques to identify influential predictors of student performance while maintaining predictive accuracy [54, 55]. These studies illustrate the broader value of interpretable predictive modelling for identifying influential variables and supporting evidence-informed educational decision-making.

4.4 Performance Evaluation Metrics

Model performance was primarily evaluated using classification accuracy, enabling comparison of predictive capability across all classifiers. In addition to predictive performance, feature importance scores were examined to identify the assessment variables most strongly associated with academic outcomes and to evaluate the relative contribution of MCQ-based and task-based assessments to final grade prediction.

5 Results

This section presents the predictive performance of the evaluated machine learning classifiers and analyses the relative importance of assessment variables in predicting final grade outcomes. Classifier accuracy results are first compared, followed by an examination of feature importance scores and their implications for the study's research questions.

5.1 Classifier Performance Comparison

Table 3 presents the classification accuracies and normalised feature importance scores obtained from the evaluated machine learning models. SVM achieved the highest classification accuracy (96.07%), followed by XGBoost (93.76%), KNN (88.45%), Extra Trees (87.76%), and Decision Tree (82.22%). Logistic Regression produced the lowest accuracy (69.75%). The superior performance of SVM and XGBoost suggests that non-linear modelling approaches were better able to capture the relationships between assessment variables and student grade outcomes.

Across the higher-performing classifiers, the Final Exam variable consistently emerged as the most influential predictor of final grade outcomes, receiving the highest importance score in the SVM, XGBoost, KNN, Extra Trees, and Decision Tree models. Assessment 3, which primarily consisted of task-based assessment activities, was also identified as one of the strongest predictors of student performance. The consistency of these findings across multiple classifiers suggests that overall examination performance and task-based assessments play a central role in forecasting academic outcomes.

In contrast, the feature importance profile produced by Logistic Regression differed substantially from the other models. Variables such as Semester and Unit

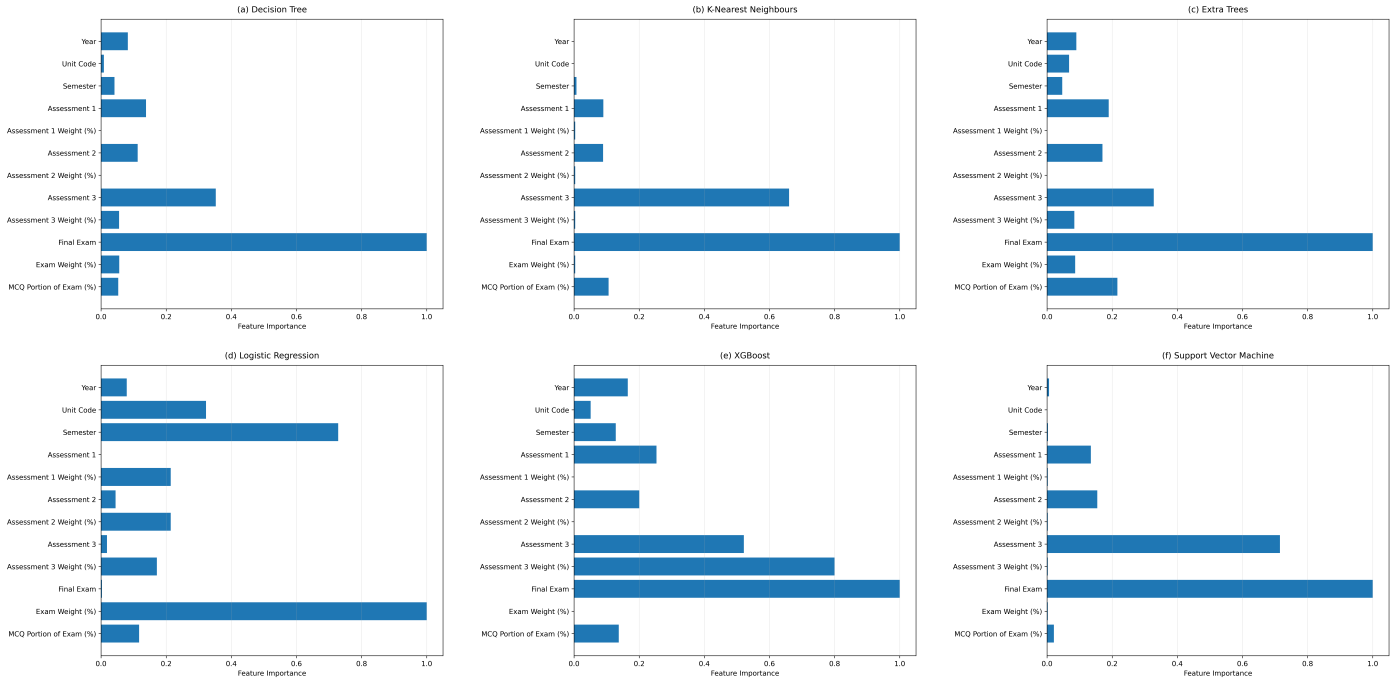


Figure 1. Comparative feature-importance profiles of the six evaluated classifiers: (a) Decision Tree, (b) K-Nearest Neighbours, (c) Extra Trees, (d) Logistic Regression, (e) XGBoost, and (f) Support Vector Machine.

Code received comparatively high importance scores despite showing limited predictive relevance in the higher-performing classifiers. Given that Logistic Regression also achieved the lowest classification accuracy (69.75%), these results suggest that the linear assumptions of the model may not adequately capture the complex relationships present within the assessment dataset. Consequently, the feature importance estimates produced by Logistic Regression should be interpreted with caution.

5.2 Feature Importance Analysis

Figure 1(a) presents the feature-importance profile of the Decision Tree classifier. Final Exam and Assessment 3 emerged as the most influential variables, indicating that final examination performance and task-based assessments were the strongest predictors of final grade outcomes within the Decision Tree model.

Across most classifiers, Assessment 1 and Assessment 2 exhibited lower importance scores than Assessment 3 and Final Exam, suggesting that online MCQ-based assessments contributed less strongly to final grade prediction than task-based assessments and overall examination performance.

Figure 1(b) presents the feature-importance profile of the K-Nearest Neighbours (KNN) classifier. Similar to the Decision Tree model, Final Exam and Assessment 3 received the highest importance scores, suggesting that these variables consistently contributed strongly to grade prediction across different modelling approaches.

Figure 1(c) presents the feature-importance profile of the Extra Trees classifier, which achieved an accuracy of 87.76%. Final Exam and Assessment 3 were the two most influential predictors, while MCQ Portion of Exam (%) received greater importance than it did in several other classifiers. This result suggests that the written-examination MCQ component contributed to prediction within the Extra Trees model, although its influence remained substantially lower than that of overall examination performance and Assessment 3. Assessments 1 and 2 also contributed to prediction, but their relative importance was lower.

As shown in Figure 1(d), Logistic Regression achieved an accuracy of 69.75%. Its feature-importance profile assigned the greatest importance to Exam Weight (%), followed by Semester and Unit Code. However, given the comparatively low classification accuracy achieved by this model, these feature-importance estimates should be interpreted cautiously. The substantially lower importance assigned to most other features further highlights the distinctive feature-importance pattern produced by Logistic Regression.

As shown in Figure 1(e), XGBoost achieved the second-highest classification accuracy, at 93.76%. Final Exam, Assessment 3 Weight (%), and Assessment 3

Table 4. Three most influential features identified by each classifier.

Method	Feature	Feature detail	Importance score
SVM	Final Exam	Written final examination containing both task-based and MCQ questions	1.000000
	Assessment 3	Task-based assessment	0.715190
	Assessment 2	Marks in online MCQ-based mid-semester Assessment 2	0.154230
XGBoost	Final Exam	Written final examination containing both task-based and MCQ questions	1.000000
	Assessment 3 Weight (%)	Weight of Assessment 3 towards the final grade	0.799870
	Assessment 3	Task-based assessment	0.521420
KNN	Final Exam	Written final examination containing both task-based and MCQ questions	1.000000
	Assessment 3	Task-based assessment	0.660380
	MCQ Portion of Exam (%)	Relative percentage of MCQs in the written examination	0.106080
Extra Trees	Final Exam	Written final examination containing both task-based and MCQ questions	1.000000
	Assessment 3	Task-based assessment	0.327630
	MCQ Portion of Exam (%)	Relative percentage of MCQs in the written examination	0.216070
DT	Final Exam	Written final examination containing both task-based and MCQ questions	1.000000
	Assessment 3	Task-based assessment	0.352420
	Assessment 1	Marks in online MCQ-based mid-semester Assessment 1	0.137990
Logistic Regression	Exam Weight (%)	Weight of the final examination towards the final grade	1.000000
	Semester	Semester in which the unit was taught	0.728090
	Unit Code	Code of the unit (subject)	0.322400

were the most influential features identified by the model, with importance scores of 1.000, 0.800, and 0.521, respectively. These results indicate that final examination performance, the weighting assigned to Assessment 3, and performance in the task-based Assessment 3 contributed strongly to grade prediction within the XGBoost model. This feature-importance pattern is consistent with XGBoost's capacity to model complex relationships and interactions among predictor variables.

As shown in Figure 1(f), SVM achieved the highest classification accuracy, at 96.07%. Final Exam and Assessment 3 emerged as the dominant predictors, with importance scores of 1.000 and 0.715, respectively, while Assessment 2 also contributed to grade prediction with an importance score of 0.154. These results further demonstrate the strong predictive contribution of final examination performance and task-based Assessment 3 to final grade outcomes.

Table 4 lists the three most influential features identified by each classifier.

Collectively, the feature importance results demonstrate a high degree of consistency across the strongest-performing classifiers. Final Exam and Assessment 3 repeatedly emerged as the most influential predictors of final grade outcomes, whereas Assessment 1 and Assessment 2 generally exhibited lower importance scores. These findings suggest that task-based assessments and overall examination performance contribute more strongly to academic success than MCQ-based assessments alone.

5.3 Research Question Analysis

5.3.1 *Impact of MCQ Components Within Examinations (RQ1)*

The feature importance analysis revealed that overall examination performance was consistently one of the strongest predictors of final grade outcomes

across all high-performing classifiers. In contrast, the variable representing the proportion of MCQ marks within the examination (MCQ Portion of Exam (%)) generally received substantially lower importance scores. This suggests that the predictive contribution of the examination is driven primarily by its broader assessment structure rather than by the MCQ component alone. Given that the final examination contained both MCQ and task-based questions, the results indicate that task-based examination components contributed more strongly to grade prediction than MCQ components. Because MCQ Portion of Exam (%) describes the design of each examination rather than students' marks on the MCQ items, this inference is indirect and should be interpreted with caution. These findings suggest that assessment tasks requiring application, analysis, and problem-solving may provide stronger indicators of overall academic achievement than questions focused primarily on knowledge recall.

5.3.2 *Online MCQ Assessments versus Task-Based Assessments (RQ2)*

Assessment 3 consistently emerged as one of the most influential predictors of student performance across nearly all classifiers. Although Assessment 1 and Assessment 2 appeared among the important variables in some models, their importance scores were generally lower than those associated with Assessment 3. Given that Assessment 3 primarily consisted of task-based assessment activities, these findings suggest that task-based assessments provide stronger predictive signals of final academic performance than online MCQ assessments.

The results indicate that assessment activities requiring students to apply knowledge, demonstrate understanding, and engage in higher-order cognitive processes contribute more strongly to grade prediction than assessments focused primarily on MCQ-based evaluation. This finding supports the value of incorporating task-based assessments alongside MCQs when evaluating overall student achievement.

5.3.3 *Online and Written MCQ Comparisons (RQ3)*

The results suggest that both online MCQ assessments (Assessments 1 and 2) and the written MCQ component of the final examination contributed to grade prediction, although their relative importance varied across classifiers. In several models, the online MCQ assessments achieved importance scores comparable to or higher than the MCQ Portion of Exam (%) variable, indicating that online MCQ performance may provide stronger predictive information than written MCQ performance alone. However, the differences were not sufficiently large or consistent across all classifiers to support a definitive conclusion regarding the superiority of one delivery mode over the other.

The observed variation may reflect differences in assessment context, timing, examination conditions, or the design of assessment items across delivery formats. Online MCQ assessments were administered throughout the semester and may therefore capture ongoing student engagement and learning progression, whereas written MCQ components formed only one part of a larger summative examination. Consequently, both assessment modes appear to provide useful information regarding student performance, although neither demonstrated the predictive strength observed for task-based assessments or overall examination performance. Further research is required to better understand how assessment delivery modes influence academic outcomes and the interpretation of learning analytics data.

6 Discussion

6.1 Interpretation of Findings

The performance of the machine learning classifiers and their associated feature importance profiles provide valuable insights into the relationship between assessment design and academic outcomes. Across almost all classifiers, Assessment 3 and final examination performance emerged as the strongest predictors of final grades, indicating the substantial predictive contribution of task-based assessments [9]. This finding is further contextualised by evidence that students tend to adopt surface-level study strategies when preparing for MCQ examinations [56], which may limit the depth of learning that MCQ-based assessments ultimately capture. These findings are consistent with research suggesting that assessments requiring application, analysis, and higher-order cognitive processing may provide stronger indicators of student achievement than assessments focused primarily on recall and recognition [8, 57].

Beyond these findings, the consistency of feature importance across multiple classifiers demonstrates the value of explainable AI for educational assessment research. Rather than treating machine learning models as black-box predictors, feature importance analysis enables educators to understand which

assessment components contribute most strongly to student performance, thereby supporting more transparent and evidence-informed curriculum and assessment design. This aligns with recent perspectives advocating interpretable and trustworthy Educational Data Mining systems that balance predictive performance with educational usefulness [1].

Although common patterns emerged across classifiers, the variation in feature importance profiles highlights the complexity of educational assessment data. Ensemble methods such as XGBoost and Extra Trees demonstrated strong predictive performance and identified a broader range of influential variables, consistent with large-scale comparative evaluations showing that ensemble classifiers consistently rank among the strongest performers across diverse real-world datasets [58], a capacity attributable to their ability to model complex non-linear interactions [51]. In contrast, Logistic Regression exhibited substantially lower predictive performance and generated feature importance patterns that differed from those produced by the higher-performing classifiers. These differences highlight the varying assumptions and sensitivities of machine learning algorithms when applied to educational datasets [35, 52].

The consistent importance assigned to overall examination performance and Assessment 3 suggests that the nature of assessment tasks plays a critical role in predicting academic outcomes. Assessment activities requiring students to apply knowledge, demonstrate understanding, and solve problems exhibited greater predictive value than MCQ-based assessments. While online MCQ assessments also contributed to grade prediction, their influence was generally weaker than that of task-based assessments. These findings support the view that balanced assessment strategies should include opportunities for both knowledge verification and deeper demonstration of learning.

Among all evaluated models, SVM achieved the highest classification accuracy (96.07%). The most influential variables within the SVM model were final examination performance, Assessment 3, and Assessment 2. The dominance of examination performance and Assessment 3 further reinforces the importance of assessment activities that evaluate students' ability to synthesise, apply, and extend knowledge. The strong performance of SVM is also consistent with previous research demonstrating its effectiveness in classification problems involving complex relationships among predictor variables [42, 47].

While the role of MCQs in the examination context appears less influential in grade prediction, it would be remiss not to consider the potential benefits of well-designed MCQs in reinforcing learning [3, 30]. The nuanced impact of MCQs in online and written formats warrants further investigation, particularly given the evolving nature of digital pedagogies and the increasing prevalence of online assessment [4, 22].

6.2 Implications for Assessment Design

These findings prompt a nuanced consideration of assessment design and highlight the distinct contributions of task-based and MCQ assessments within both formative and summative learning environments. The results suggest that while MCQs remain valuable for efficiently assessing foundational knowledge, task-based assessments appear to provide stronger indicators of overall student achievement and higher-order cognitive performance. From a curriculum design perspective, this supports the use of balanced assessment strategies that combine efficient knowledge testing with opportunities for application, analysis, and problem-solving. Educational practitioners may therefore benefit from carefully aligning assessment formats with intended learning outcomes, ensuring that assessment practices measure not only knowledge recall but also deeper learning capabilities.

7 Explainability, Fairness, and Ethical Considerations

As artificial intelligence techniques become increasingly integrated into educational decision-making and learning analytics systems, issues of transparency, fairness, and interpretability become increasingly important. Although the objective of this study was not to make automated decisions regarding individual students, the findings demonstrate how machine learning models can identify assessment variables that are strongly associated with academic outcomes.

Feature importance analysis was therefore included to improve model interpretability and provide educators with insight into the factors driving predictions. This aligns with recent recommendations advocating Educational Data Mining systems that extend beyond predictive accuracy by incorporating transparency, interpretability, fairness, and

accountability as fundamental design principles [1]. Such characteristics are particularly important when AI models may influence educational policy, assessment practices, or student support decisions. Nevertheless, caution should be exercised when using predictive models in educational settings, as historical assessment data may contain biases associated with assessment design, student demographics, or contextual learning factors. Furthermore, student privacy and responsible use of educational data remain important considerations when developing AI-driven learning analytics systems. Consequently, predictive models should be viewed as decision-support tools rather than replacements for academic judgement and pedagogical expertise.

8 Conclusion

This study investigated the extent to which multiple-choice question (MCQ) assessments reflect student performance by applying machine learning classification techniques to assessment data collected from 2,390 students enrolled across sixteen offerings of three undergraduate and postgraduate economics units at Charles Darwin University between 2016 and 2019. Drawing on a dataset comprising 26,698 non-missing attribute values, the study examined the relative contribution of online MCQ assessments, written examination MCQ components, task-based assessments, and final examination performance in predicting final grade outcomes.

Among the evaluated machine learning models, Support Vector Machines (SVM) achieved the highest classification accuracy (96.07%), followed by XGBoost (93.76%), indicating the suitability of non-linear classification approaches for modelling assessment-performance relationships in higher education datasets. Feature importance analysis consistently identified Assessment 3 and final examination performance as the strongest predictors of student grades across most classifiers. In contrast, the MCQ component of examinations generally exhibited lower predictive importance, suggesting that task-based assessments provide stronger indicators of overall academic achievement and higher-order cognitive performance.

The findings demonstrate that while MCQs remain valuable tools for efficiently assessing foundational knowledge and supporting formative learning, assessments requiring application, analysis, and problem-solving contribute more strongly to predicting academic success. These results support the adoption of balanced assessment strategies that combine the efficiency of MCQs with opportunities for students to demonstrate deeper understanding through task-based activities.

From an Educational Data Mining perspective, this research illustrates how machine learning and explainable AI techniques can be used to evaluate the relative contribution of assessment components and provide evidence-based insights into assessment design. More broadly, the findings support emerging directions in Educational Data Mining that advocate AI systems that are not only accurate but also transparent, interpretable, and capable of supporting trustworthy educational decision-making [1].

Several limitations should be acknowledged. The dataset was drawn from a single institution and a specific disciplinary context, which may limit the generalisability of the findings. In addition, because final grades are computed from the weighted assessment components that also serve as model inputs, the classification task partly reconstructs the grading scheme; the feature importance scores therefore reflect both predictive association and the assessment weightings, and should be interpreted accordingly. Future research could extend this work by incorporating data from multiple institutions and disciplines, applying additional validation procedures and evaluation metrics, and exploring more advanced explainable AI approaches to further understand the relationship between assessment design and student success while supporting increasingly transparent and evidence-informed educational practice.

Data Availability Statement

The de-identified data supporting the findings of this study are available from the corresponding author upon reasonable request, subject to institutional ethics approval and privacy requirements.

Funding

This work was supported without any funding.

Conflicts of Interest

Darius Pfitzner is affiliated with the Australian Institute of Business, Adelaide, SA 5000, Australia. The authors declare that this affiliation had no influence on the study design, data collection, analysis, interpretation, or the decision to publish, and that no other competing interests exist.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

The Charles Darwin University Human Research Ethics Committee determined that the secondary analysis of the de-identified records was exempt from further ethical review in accordance with institutional guidelines for research involving de-identified data.

References

[1] Keivanian, F., & Fan, Z. (2026). Beyond accuracy: Toward interpretable, multi-objective, and trustworthy educational data mining systems. *ICCK Transactions on Educational Data Mining, 2*(2), 52-55. [CrossRef]

[2] Nair, G. G., & Feroze, M. (2023). Effectiveness of multiple-choice questions (MCQs) discussion as a learning enhancer in conventional lecture class of undergraduate medical students. *Medical Journal of Dr. D.Y. Patil Vidyapeeth, 16*(Suppl 2), S183-S188. [CrossRef]

[3] Manoharan, S. (2019). Cheat-resistant multiple-choice examinations using personalization. *Computers & Education, 130*, 139-151. [CrossRef]

[4] Williams, J. B., & Wong, A. (2009). The efficacy of final examinations: A comparative study of closed-book, invigilated exams and open-book, open-web exams. *British Journal of Educational Technology, 40*(2), 227-236. [CrossRef]

[5] Liu, Q., Wald, N., Daskon, C., & Harland, T. (2024). Multiple-choice questions (MCQs) for higher-order cognition: Perspectives of university teachers. *Innovations in Education and Teaching International, 61*(4), 802-814. [CrossRef]

[6] Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition, 7*(3), 323-331. [CrossRef]

[7] Joorabchi, B., & Chawhan, A. R. (1975). Multiple choice questions. The debate goes on. *British Journal of Medical Education, 9*(4), 275-280. [CrossRef]

[8] Brown, G. T., & Abdulnabi, H. H. (2017, June). Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. In *Frontiers in Education* (Vol. 2, p. 24). Frontiers Media SA. [CrossRef]

[9] Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? *BMC medical education, 7*(1), 49. [CrossRef]

[10] Koretsky, M. D., Brooks, B. J., & Higgins, A. Z. (2016). Written justifications to multiple-choice concept questions during active learning in class. *International Journal of Science Education, 38*(11), 1747-1765. [CrossRef]

[11] Bhatt, M., & Dua, S. (2016). Use of multiple choice questions during lectures helps medical students improve their performance in written formative assessment in physiology. *National Journal of Physiology, Pharmacy and Pharmacology, 6*(6), 576. [CrossRef]

[12] Hauer, K. E., Boscardin, C., Brenner, J. M., van Schaik, S. M., & Papp, K. K. (2020). Twelve tips for assessing medical knowledge with open-ended questions: Designing constructed response examinations in medical education. *Medical Teacher, 42*(8), 880-885. [CrossRef]

[13] Kang, S. H., McDermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European journal of cognitive psychology, 19*(4-5), 528-558. [CrossRef]

[14] Kuechler, W. L., & Simkin, M. G. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovative Education, 8*(1), 55-73. [CrossRef]

[15] Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*(1), 77-92. [CrossRef]

[16] Kennedy, P., & Walstad, W. B. (1997). Combining multiple-choice and constructed-response test scores: An economist's view. *Applied measurement in education, 10*(4), 359-375. [CrossRef]

[17] Becker, W. E., & Johnston, C. (1999). The relationship between multiple choice and essay response questions in assessing economics understanding. *Economic Record, 75*(4), 348-357. [CrossRef]

[18] Dufresne, R. J., Leonard, W. J., & Gerace, W. J. (2002). Making sense of students' answers to multiple-choice questions. *The Physics Teacher, 40*(3), 174-180. [CrossRef]

[19] Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163-184. [CrossRef]

[20] Bognár, L., Fauszt, T., & Váraljai, M. (2021). The impact of online quizzes on student success. *International Journal of Emerging Technologies in Learning (iJET), 16*(11), 225-244. [CrossRef]

[21] Massoudi, D., Koh, S., Hancock, P. J., & Fung, L. (2017). The effectiveness of usage of online multiple choice questions on student performance in introductory accounting. *Issues in Accounting Education, 32*(4), 1-17. [CrossRef]

[22] Blondeel, E., Everaert, P., & Opdecam, E. (2024). Does practice make perfect? The effect of online formative assessments on students' self-efficacy and

test anxiety. *The British Accounting Review, 56*(4), 101189. [CrossRef]

[23] Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The internet and higher education, 27*, 1-13. [CrossRef]

[24] Xu, D., & Jaggars, S. S. (2014). Performance gaps between online and face-to-face courses: Differences across types of students and academic subject areas. *The Journal of Higher Education, 85*(5), 633-659. [CrossRef]

[25] Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 14*(2), 194-199. [CrossRef]

[26] Zakay, D., & Glicksohn, J. (1992). Overconfidence in a multiple-choice test and its relationship to achievement. *The Psychological Record, 42*(4), 519-524. [CrossRef]

[27] Funk, S. C., & Dickson, K. L. (2011). Multiple-choice and short-answer exam performance in a college classroom. *Teaching of Psychology, 38*(4), 273-277. [CrossRef]

[28] Rovai, A., Ponton, M., Wighting, M., & Baker, J. (2007, July). A comparative analysis of student motivation in traditional classroom and e-learning courses. In *International Journal on E-learning* (Vol. 6, No. 3, pp. 413-432). Association for the Advancement of Computing in Education (AACE). [CrossRef]

[29] Broadbent, J. (2017). Comparing online and blended learner's self-regulated learning strategies and academic performance. *The Internet and Higher Education, 33*, 24-32. [CrossRef]

[30] Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604-616. [CrossRef]

[31] Nicol, D. (2007). E-assessment by design: Using multiple-choice tests to good effect. *Journal of Further and Higher Education, 31*(1), 53-64. [CrossRef]

[32] Wilkie, R. M., Harley, C., & Morrison, C. (2009). High level multiple choice questions in advanced psychology modules. *Psychology Learning & Teaching, 8*(2), 30-36. [CrossRef]

[33] Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education, 35*(4), 453-472. [CrossRef]

[34] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260. [CrossRef]

[35] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). [CrossRef]

[36] Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & education, 143*, 103676. [CrossRef]

[37] Hu, J. (2021). Teaching evaluation system by use of machine learning and artificial intelligence methods. *International Journal of Emerging Technologies in Learning (iJET), 16*(5), 87-101. [CrossRef]

[38] Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer. https://link.springer.com/book/9780387310732

[39] Mahesh, B. (2020). Machine learning algorithms - A review. *International Journal of Science and Research (IJSR), 9*(1), 381-386. [CrossRef]

[40] Balaji, P., Alelyani, S., Qahmash, A., & Mohana, M. (2021). Contributions of machine learning models towards student academic performance prediction: A systematic review. *Applied Sciences, 11*(21), 10007. [CrossRef]

[41] Sen, P. C., Hajra, M., & Ghosh, M. (2019). Supervised classification algorithms in machine learning: A survey and review. In *Emerging technology in modelling and graphics: Proceedings of IEM graph 2018* (pp. 99-111). Singapore: Springer Singapore. [CrossRef]

[42] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing, 408*, 189-215. [CrossRef]

[43] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery, 10*(3), e1355. [CrossRef]

[44] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297. [CrossRef]

[45] Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press. [CrossRef]

[46] Schölkopf, B., & Smola, A. J. (2002). Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT Press. [CrossRef]

[47] Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision analytics journal, 3*, 100071. [CrossRef]

[48] Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine learning, 6*(1), 37-66. [CrossRef]

[49] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory,*

13(1), 21-27. [CrossRef]

[50] Priyanka, & Kumar, D. (2020). Decision tree classifier: a detailed survey. *International Journal of Information and Decision Sciences, 12*(3), 246-269. [CrossRef]

[51] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning, 63*(1), 3-42. [CrossRef]

[52] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley. [CrossRef]

[53] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: with applications in R* (Vol. 2). New York: springer. [CrossRef]

[54] Alamri, R., & Alharbi, B. (2021). Explainable student performance prediction models: A systematic review. *IEEE Access, 9*, 33132-33143. [CrossRef]

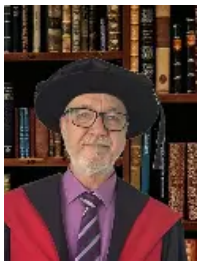
[55] Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., ... & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and education: artificial intelligence, 3*, 100074. [CrossRef]

[56] Scouller, K. M., & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education, 19*(3), 267-279. [CrossRef]

[57] Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research, and Evaluation, 22*, 1-13. [CrossRef]

[58] Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *The journal of machine learning research, 15*(1), 3133-3181. https://www.jmlr.org/papers/volume15/delgado14a/delgado14a.pdf

Niusha Shafiabady is Professor and Head of the Discipline of Information Technology at Australian Catholic University and Director of the Women in AI for Social Good Lab. Her research interests include artificial intelligence, computational intelligence, evolutionary optimisation, explainable AI, machine learning, data analytics, computational optimisation, and ethical AI. Her research emphasises the development of trustworthy and socially responsible AI technologies for real-world applications. She has led interdisciplinary research projects and contributed to research in AI for social good, digital innovation, and sustainable development. (Email: Niusha.Shafiabady@acu.edu.au)



Darius Pfitzner is the Academic Dean at the Australian Institute of Business. He has more than two decades of experience in higher education and has held senior academic leadership roles across Australian universities, including Associate Dean positions in learning, teaching, and innovation. His research and professional interests include information technology, digital education, curriculum design, educational leadership, and student-centred learning. He also has extensive experience in academic quality assurance, professional accreditation, and the development of innovative teaching practices aligned with industry needs. (Email: Darius.Pfitzner@aib.edu.au)

Talha Nawaz is an artificial intelligence researcher whose research interests include explainable artificial intelligence, computer vision, deep learning, and machine learning. His research focuses on developing intelligent systems for real-world applications, including computer vision and AI-driven data analysis. He has contributed to research on explainable AI and artificial intelligence applications across multidisciplinary domains, including healthcare and media analytics. (Email: talha.nawaz@fau.de)

Seb Dianati is an Associate Professor and Senior Academic Development Consultant in Learning and Teaching Innovation at the University of Southern Queensland. His research interests include educational technology, digital learning, generative artificial intelligence in higher education, student partnerships, curriculum innovation, and technology-enhanced learning. He has led and contributed to projects that advance inclusive, student-centred learning and the effective integration of digital technologies into higher education. His work has been published in education journals and conference proceedings. (Email: Seb.Dianati@unisq.edu.au)

Mark Brady is a Senior Lecturer in Law at Charles Darwin University. His research interests include law and emerging technologies, artificial intelligence governance, automated vehicles, legal ethics, digital regulation, healthcare law, and technology policy. He has published on the legal and ethical implications of artificial intelligence, automated vehicles, privacy, and commercial surrogacy, with particular emphasis on developing legal frameworks that support responsible technological innovation. (Email: mark.brady@cdu.edu.au)