Sains Insani eISSN: [0127-7871]

Beyond Grades: Predicting Programme Learning Outcomes with Multi-Output Regression in Malaysian Higher Education

Melangkaui Gred: Meramal Hasil Pembelajaran Program (PLO) dengan Regresi Pelbagai Hasil dalam Pendidikan Tinggi Malaysia

Muhamad Noorazizi Abd Ghani[1] **& Istas Fahrurrazi Nusyirwan**[2]

[1] *Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia*
[2] *Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia*

Article progress
Received: 3 November 2025
Accepted: 24 November 2025
Published: 31 May 2025

*Corresponding author:
Istas Fahrurrazi Nusyirwan,
Malaysia-Japan International Institute of Technology,
Universiti Teknologi Malaysia,
Kuala Lumpur, Malaysia;
Email: istaz@utm.my

Abstract: Grades often conceal specific competencies, while Malaysia's Outcome-Based Education (OBE) requires evidence at the Programme Learning Outcome (PLO) level. We develop a PLO-centric predictive framework that forecasts students' 12 end-of-programme PLO scores. Using anonymised records for 194 engineering students, we constructed a semester-indexed feature set (167 predictors) and jointly modelled all 12 PLOs via multi-output regression. Eleven algorithms (CatBoost, Extra Trees, LightGBM, Gradient Boosting, XGBoost, Random Forest, SVR, k-NN, Bayesian Ridge, AdaBoost, HistGradientBoosting) were evaluated under 70/30, 80/20, 90/10 hold-out regimes with 10-fold CV on training folds. We benchmarked macro Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of determination (R^2), and tested stacking and convex blending. Across splits, Bayesian Ridge was the most reliable single model by MAE (with CatBoost occasionally leading in R^2), and a simple convex blend (Bayesian Ridge + CatBoost ± Extra Trees/Gradient Boosting) delivered ~5-13% lower MAE than the best single learner, whereas stacking did not add consistent gains. Model interpretability (model-native importances, SHapley Additive exPlanations (SHAP)) shows a noticeable later-year (third year and beyond) pattern, Semester 2 signal with recurrent contributions from PLO2 and PLO7, alongside a meaningful early-year indicator (e.g., PLO2_Semester_2_Year_1). The framework enables earlier, PLO-level feedback aligned with accreditation, supporting targeted intervention and curriculum improvement can be extended with broader cohorts, richer behavioural/demographic features, fairness audits, and cross-institutional generalisation.

Keywords: Programme Learning Outcomes; multi-output regression; outcome-based education; Malaysia higher education

Abstrak: Gred sering menyembunyikan kecekapan khusus, manakala Pendidikan Berasaskan Hasil (Outcome-Based Education, OBE) di Malaysia memerlukan bukti pada aras Hasil Pembelajaran Program (Program Learning Outcome, PLO). Kami membangunkan rangka kerja ramalan berpusatkan PLO yang meramalkan 12 skor PLO akhir program pelajar. Menggunakan rekod yang telah dianonimkan bagi 194 pelajar kejuruteraan, kami membina set ciri berindeks semester (167 peramal) dan memodelkan kesemua 12 PLO secara bersama melalui regresi pelbagai hasil. Sebelas algoritma (CatBoost, Extra Trees, LightGBM, Gradient Boosting, XGBoost, Random Forest, SVR, k-Jiran Terdekat (k-NN), Bayesian Ridge, AdaBoost, HistGradientBoosting) dinilai di bawah tiga pembahagian latih-ujian jenis hold-out (70/30, 80/20, 90/10) dengan pengesahan silang 10 lipatan pada set latihan. Kami membanding aras ralat menggunakan ralat mutlak min berpurata makro (MAE), ralat punca min kuasa dua (RMSE) dan pekali penentuan (R^2), serta menguji dua strategi ensambel (penindanan dan pengadunan cembung). Merentas semua pecahan, Bayesian Ridge ialah model tunggal paling boleh diharap dari segi MAE (dengan CatBoost kadang-kadang mendahului dari segi R^2), dan pengadunan

PUSAT PENGAJIAN TERAS
Centre of Core Studies
USIM Universiti Sains Islam Malaysia

cembung ringkas (Bayesian Ridge + CatBoost ± Extra Trees/Gradient Boosting) memberikan pengurangan MAE sekitar ~5-13% berbanding model tunggal terbaik, manakala penindanan tidak menambah peningkatan yang konsisten. Kebolehtafsiran model (kepentingan asli model dan SHapley Additive exPlanations (SHAP)) menunjukkan corak yang jelas pada tahun-tahun akhir (tahun ketiga dan ke atas), isyarat Semester 2 dengan sumbangan berulang daripada PLO2 dan PLO7, di samping penunjuk awal tahun yang bermakna (cth., PLO2_Semester_2_Tahun_1). Rangka kerja ini membolehkan maklum balas lebih awal pada aras PLO yang sejajar dengan keperluan akreditasi, menyokong intervensi tersasar dan penambahbaikan kurikulum. Pada masa hadapan, ia boleh diperluas dengan kohort yang lebih besar, ciri tingkah laku/demografi yang lebih kaya, audit keadilan, serta penggeneralisasian rentas institusi.

Kata kunci: Hasil Pembelajaran Program (PLO); regresi pelbagai hasil; Pendidikan Berasaskan Hasil (OBE); pendidikan tinggi Malaysia

Introduction

Predictive models built on grades or Grade Point Average (GPA) may inherit the limits of grade-centred assessment: susceptibility to inflation, mixed constructs, and weak alignment to specific competencies, yielding coarse feedback for teaching and support (Almaghaslah, 2025; Jaehn et al., 2025; Khan et al., 2025; Mio, 2024). In outcome-based settings, this coarseness is problematic because decisions should reflect what students know and can do.

In Malaysia, engineering degree programmes are accredited under an OBE framework, and programmes must assess and evidence attainment of PLOs for accreditation and continuous improvement (Pao Liew & Loo Kiew, 2022). For engineering programmes, this shifts emphasis from aggregate marks to demonstrable attainment across defined competencies, creating demand for analytics that operate at the PLO level rather than the GPA level.

Recent work in learning analytics still leans toward aggregate outcomes (e.g., grades/pass-fail), while PLO-level prediction remains relatively rare and typically categorical rather than continuous. This study develops a PLO-centric predictive framework that forecasts twelve end-of-programme PLO scores from per-semester attainment across each student's entire enrolment duration (often eight semesters, occasionally more) using multi-output regression. We evaluate eleven algorithms, explore stacking and convex blending, and use model native importance and SHAP for explanation. Across train-test regimes, Bayesian Ridge delivered the lowest MAE, and a simple convex blend cut MAE further; we also summarise which early-semester PLO signals consistently drive predictions. We pose three research questions (RQs):

Research Questions

- **RQ1**. To what extent can multi-output regression accurately predict the twelve end-of-programme PLO scores from semester-wise attainment signals?
- **RQ2**. Which algorithms (tree/boosting, linear, kernel, instance-based) perform best across 70/30, 80/20 and 90/10 splits under MAE, RMSE and R^2?
- **RQ3**. Which predictors most strongly drive the predictions, based on model-native importances and SHAP?

Related Work

Predictive analytics has become a core practice in learning analytics and educational data mining. Systematic reviews through the early 2020s show a steady increase in the use of machine learning and deep learning models to forecast academic outcomes (Ngulub & Masumbika Ncube, 2025; Sghir et al., 2023). The number of studies accelerated after 2012 due to growing access to digital trace data and better data processing tools. Early efforts focused on relatively simple prediction tasks, such as forecasting final grades or identifying dropout risk, using aggregated academic indicators like course grades, cumulative GPA (CGPA), and demographic attributes

(Czerkawski, 2015; Olaleye & Vincent, 2020). These problems were typically framed as binary or multi-class classification challenges, with early models built using algorithms such as support vector machines (SVM), decision trees (DT), and k-nearest neighbours (k-NN) (Asiah et al., 2019; Pali & Verma, 2024). Over the past decade, the field has shifted toward more sophisticated approaches. Ensemble methods, such as random forests (RF), gradient boosting, and stacking, now routinely outperform traditional models by capturing nonlinear relationships and improving robustness in predictive tasks (Dong et al., 2020; Rane et al., 2024).

Early predictive analytics research also concentrated on classifying students, such as pass/fail or various levels of performance. These studies typically relied on aggregated academic indicators: course marks or CGPA, and employed traditional algorithms like logistic regression, naïve Bayes (NB), k-NN, DT, and SVM (Ab Rahim & Buniyamin, 2022; Jović et al., 2022). Beyond early classifiers, recent work explores continuous prediction and ensemble methods, often with explainability. While regression models provide an alternative approach by predicting continuous measures of achievement, such as GPA. One study partitioned predictors into psychological, sociological and study-related factors and evaluated linear regression, DT and RF models; the RF model achieved the lowest mean absolute percentage error (~11.1 %), highlighting the utility of ensemble trees for handling nonlinear relationships (Falát & Piscová, 2022).

As machine learning techniques matured, researchers began to combine multiple models to improve prediction accuracy and capture diverse patterns. Recent work by Abiodun and Wreford (2024) demonstrated a hybrid ensemble model combining RF, k-NN, and XGBoost, achieving a high R^2 score of 0.9705 in student performance prediction. Hybrid approaches increasingly integrate unsupervised methods or explainable AI tools into the modelling pipeline. For instance, Ghimire et al. (2024) applied a hybrid support vector regression model optimised with Tree-structured Parzen Estimation (TPE), combined with SHAP and Local Interpretable Model Agnostic Explanation (LIME) explainability tools, to predict final scores in a tertiary science course, reporting superior accuracy and interpretable feature contributions. Similarly, Raji et al. (2024) used a stacked ensemble (XGBoost, ExtraTrees, RF) to predict the academic performance of deaf scholars, achieving 92.99% accuracy while using SHAP and LIME to clarify risk factors. These studies demonstrate that ensemble and hybrid models, particularly when augmented with explainable AI (xAI), not only improve predictive accuracy but also provide transparency, an essential factor for trust and decision-making in educational environments. Model choice interacts with what is measured.

Predictive models in education typically rely on a combination of student-related features drawn from several categories. A recent systematic review categorises these inputs into prior academic records, demographic data, academic achievement, behavioural logs, and psychological measures (Sghir et al., 2023). Academic performance indicators like course grades and exam scores remain prevalent in most predictive models (Alturki et al., 2022). Demographic attributes, such as family income, parents' qualifications and interaction with teachers, are also influential (Kamal & Ahuja, 2019). Prior academic data, like high school final marks, are particularly common in early-stage predictions for dropout (Zanellati et al., 2024). Despite increasing interest, socio-emotional and motivational indicators are still underutilised, although recent work highlights their value in improving prediction accuracy (Rafiq et al., 2025).

In terms of target variables, the majority of predictive analytics research focuses on aggregated metrics such as course grades, semester GPAs, or CGPAs. Systematic reviews and empirical studies confirm that final exam results and overall GPA remain the most frequently predicted outcomes in educational models (Falát & Piscová, 2022; Raji et al., 2024; Zhu, 2024). While these measures offer a quantifiable benchmark, they are prone to grade inflation, which undermines their reliability as indicators of true academic achievement (Park & Cho, 2023). Moreover, such GPA-based models typically fail to reflect mastery of specific learning outcomes or competencies, such as critical thinking or communication skills. To address these gaps, researchers have turned to xAI techniques like SHAP and LIME to enhance transparency. For example, E. Ben George et al. (2025) used RF and xAI to highlight the influence of midterms and practical scores on final grades, while El Jihaoui et al. (2025) showed how poverty and class size affected CGPA predictions. These studies emphasise the importance of incorporating diverse features and interpretable methods to provide more equitable and actionable insights. Despite their ubiquity, grades/CGPA can obscure competency attainment and are susceptible to inflation, which weakens actionability. This has prompted a shift toward outcomes-aligned targets and interpretable models, but adoption remains uneven. Because grades provide coarse signals, researchers increasingly turn to learning outcomes data.

Learning outcomes, both Course Learning Outcomes (CLOs) and PLOs, define the knowledge, skills, and values students are expected to achieve and are central to accreditation frameworks such as Accreditation Board for Engineering and Technology (ABET) and OBE systems. Compared to GPA or course grades, PLOs offer more fine-grained indicators of competency attainment. However, they remain underutilised in predictive analytics, largely due to difficulties in data collection and system integration complexity (Agha et al., 2023; Zayani et al., 2024). Earlier studies often used GPA or grades as proxy targets due to their wide availability, while studies attempting to predict PLOs were often based on simple models like linear regression (Nor Afiqah Wan Othman et al., 2020). Within OBE, course-level prediction using mid-semester signals shows practical promise: across 30 courses and 2,423 students, Tjandra et al. (2024) compared DT, RF, Support Vector Regressor (SVR) and k-NN, finding SVR strongest in large classes (≥100; MAE≈0.06, RMSE≈0.07, R^2≈0.90 in Discrete Mathematics), while tree ensembles fared better in smaller cohorts. These developments highlight a growing but still early-stage shift toward using learning outcomes as both pedagogical and predictive instruments. These efforts indicate momentum toward outcomes-aligned prediction, yet programme-level PLO forecasting with continuous targets is still rare, and most studies remain single-output or course-specific.

The field of learning analytics has only recently begun to exploit multi-output regression in educational applications. Xue and Niu's hybrid ensemble model are an example of a multi-output hybrid ensemble applied to higher education data, using data from an online learning platform; they jointly predicted homework, midterm, experiment, and final exam scores. When XGBoost was trained on the first six weeks of behavioural data, it achieved 78.37% accuracy for midterm and final grades, 3- 8 percentage points higher than comparison models, many of which are single-output; gradient boosting produced a mean squared error of 16.76 for homework and experiment grades (Xue & Niu, 2023). The authors argued that multi-output prediction provides earlier and more granular insights than conventional single-grade models. Another line of work treats related classification outputs jointly: Yekun and Haile (2021) developed a multi-label ensemble model that predicts next semester performance across five high school courses, demonstrating that label powerset transformations and ensemble methods outperform binary relevance and classifier chain baselines. Although this work focuses on discrete labels rather than continuous scores, it suggests that simultaneous prediction of multiple educational outcomes can improve accuracy and enable early interventions.

Multi-output regression, which predicts two or more (often continuous) outcomes simultaneously, has seen growing use in other technical domains such as vegetation quality, vehicle wind noise spectrum modelling, stock selection, and remote-sensing biophysical parameter estimation (Wang et al., 2024; Zhang et al., 2023), yet its application in educational contexts, particularly using PLOs as continuous targets, remains sparse. Taken together, these strands expose a specific gap.

Prior work in educational prediction has largely centred on aggregated outcomes (grades/CGPA), with recent advances in ensembles and xAI improving performance and interpretability. Studies that engage learning outcomes data (CLO/PLO) are fewer, often course-level, and typically single-output or small-sample. Very little research models programme-level PLOs as simultaneous continuous targets, limiting actionable, accreditation-aligned insight. We address this gap by evaluating multi-output regression that jointly forecasts twelve PLOs from semester-indexed attainment, comparing modern ensembles and analysing explanation via model-native importances and SHAP.

Methodology

This study adopts a retrospective predictive modelling design to forecast students' attainment of the twelve Programme Learning Outcomes (PLO1-PLO12) at graduation. We treat the end-of-programme PLO scores as continuous targets and learn from semester-wise PLO attainment profiles across semesters for cohorts entering between 2015 and 2019. This is a retrospective observational study using anonymised secondary OBE records; no interventions were applied. The scope and descriptors of the twelve PLO domains used in this study are adapted from the Faculty's Undergraduate Handbook (Session 2019/2020; see Table 1).

Table 1. Twelve Programme Learning Outcomes (PLOs) and their descriptors.

Programme Learning Outcomes	Attributes
PLO1	Engineering Knowledge
PLO2	Problem Analysis
PLO3	Design / Development of Solutions
PLO4	Investigation
PLO5	Modern Tools Usage
PLO6	The Engineer and Society
PLO7	Environment and Sustainability
PLO8	Ethics
PLO9	Communication
PLO10	Team Working
PLO11	Lifelong Learning
PLO12	Project Management, Finance & Entrepreneurship

Note. PLO descriptors adapted from Faculty of Engineering, Undergraduate Handbook, Session 2019/2020.

Raw course-level records were exported from the institutional OBE database. Each record contained a student identifier, course information (e.g., course code, semester, session), instructor information, and attainment levels for learning outcomes. During data preparation, students were restricted to one engineering programme and to academic sessions 2015/2016 through 2019/2020. The initial extract contained 484 unique students before applying the local-only filter, where locally enrolled students were retained using a 12-digit MyKad (Malaysia's government-issued national identity card) filter. After filtering and anonymisation, the final analysis cohort comprised 206 unique students. To protect privacy, all direct identifiers were removed, and the matriculation number was replaced with deterministic programme-scoped synthetic identifications (IDs). A one-to-one mapping file was stored separately under restricted access. Schema checks confirmed that the anonymised and original tables were identical except for the identifier column, ensuring no drift or loss of information.

Next is creating a modelling matrix; the aggregated student-semester table was reshaped to a wide format. We constructed a "Semester_Label" by concatenating the semester number and the year (e.g., Semester_1_Year_1). The data were reshaped to long format (Matric_ID, Semester_Label, PLO, Score) and then reshaped to a student-level wide matrix with one column per PLO-semester pair (e.g., PLO10_Semester_2_Year_1). A total of 192 semester-indexed predictors were generated. The outcomes, "PLO1_Final" to "PLO12_Final", were computed as the mean of available semester scores for each PLO per student. Merging these targets with the semester-indexed features yielded a modelling matrix of 206 students × 205 columns (192 predictors + 12 targets + 1 ID).

All PLO1-PLO12 fields were coerced to numeric and audited against the 0-100 scale. Twenty-four out-of-range cells were found across five PLOs; these were logged for traceability and capped at 100 before any aggregation or modelling. PLO13-PLO15 were not part of the mapped programme outcomes and were dropped. Course level records were aggregated to the student-semester level by computing a mean that excludes zero values, where, to facilitate data processing, unobserved missingness was temporarily flagged with 0 and then restored to Not a Number (NaN) before any statistical analysis or modelling.

Before modelling, we required sufficient longitudinal coverage by keeping only students who had reached at least their third year, evidenced by at least one recorded end-of-programme PLO value in that period. A manual audit showed that one- to two-year end-of-program PLO records were usually non-completers; accordingly, third-year progression served as a practical proxy for completion progress. This criterion retained 199 of 206 students (7 removed), with removed IDs logged in an audit workbook. 5 students had missing final PLO targets and were excluded, yielding a final modelling cohort of 194 students.

To evaluate generalisation and minimise overfitting, we applied three hold-out regimes to the final analytic cohort (n = 194) obtained after requiring third year-and-beyond PLO coverage and complete targets. The splits were 70/30 (train = 135, test = 59), 80/20 (train = 155, test = 39), and 90/10 (train = 174, test = 20), providing a robustness check across progressively smaller test sets. Within each regime, model tuning used 10-fold cross-validation on the training portion, with the test set held out once for final evaluation.

The splits were generated before any model fitting or preprocessing. All NaN predictors were identified only on the training partition and dropped from both the training and test sets to avoid leakage. Numerical features were processed through a scikit learn Pipeline consisting of median imputation (via SimpleImputer(strategy="median")) followed by min-max scaling (MinMaxScaler()), mapping each feature to the [0,1] interval. Only predictors were imputed and min-max scaled (fit on training data/folds); the twelve targets ("PLO1_Final"-"PLO12_Final") remained on their native 0-100 scale, and all evaluation metrics (MAE, RMSE, R^2) were computed on that original scale. The imputer and scaler were fit exclusively on the training data and then applied to the test data and cross-validation folds. The fitted preprocessing pipeline was saved for reproducibility.

The forecasting task was formulated as a multi output regression. Each algorithm was wrapped in MultiOutputRegressor when a native multi-target implementation was unavailable. In this study, we use independent multi-output regression via MultiOutputRegressor, training one model per PLO and not explicitly modelling inter-PLO dependencies (i.e., targets are predicted independently, conditioned only on the shared features). The model suite comprised:

- Tree/Boosting Ensembles: CatBoost, Extra Trees, Gradient Boosting, HistGradientBoosting, LightGBM, Random Forest and XGBoost.
- Linear and Kernel Models: Bayesian Ridge, Support Vector Regression (SVR), K Nearest Neighbours (KNN) and AdaBoost.

All models were initialised with default or mildly tuned hyperparameters (such as random_state=42 for reproducibility). Predictions were clipped to the [0,100] range to respect the score scale.

Performance was reported using macro-averaged metrics across the 12 PLO targets:

- MAE - the primary measure of predictive accuracy.
- RMSE - the square root of mean squared error, sensitive to large deviations.
- R^2 - the proportion of variance in the outcomes explained by the model.

Within each train/test split, we ran 10-fold shuffled cross-validation on the training portion, refitting the preprocessing pipeline (imputer and scaler) inside each fold to prevent leakage. For every model and fold, we computed MAE, RMSE, and R^2 for each of the twelve PLO targets; we also report a macro score defined as the unweighted mean across targets. To compare models, we applied a Friedman test to cross-validated errors; when the null was rejected, we conducted Nemenyi post-hoc comparisons to locate pairwise differences, and used paired Wilcoxon signed-rank tests for targeted contrasts with the selected champion.

To explore whether combining top models could yield further gains, two ensemble strategies were evaluated:

- **Stacked Generalisation (Stacking)**: Out-of-fold predictions from selected base learners (for example, CatBoost, Extra Trees and Gradient Boosting) were used to train a Ridge regression meta learner. This stacking model was then used to generate predictions on the test set.
- **Convex Blending**: A coarse grid search over non-negative weights (summing to one) was performed on out-of-fold predictions of the top models to minimise macro-MAE. The optimal weight triplets were applied to combine model predictions at test time.

To interpret the predictors driving PLO forecasts, model native feature importance and SHAP values were computed for the best single models and for the convex blends in each split; we then ranked features within each split and summarised cross-split recurrence and mean normalised importance.

Analyses were conducted in Python 3.13 with packages including Pandas 2.2, NumPy 2.1, Scikit learn 1.6, CatBoost 1.2, XGBoost 3.0, LightGBM 4.6 and Joblib 1.4. All experiments were executed on a 64-bit Windows 11 system (Core i7 CPU, 32 GB RAM, NVIDIA RTX 4070 GPU), and random seeds were fixed for reproducibility. Artefacts such as fitted preprocessors, per-fold metrics and ensemble weights were saved with timestamped filenames to ensure that results can be regenerated and audited.

Results

The final modelling dataset comprised 194 students obtained after requiring third year-and-beyond PLO coverage and complete targets. We detected 192 semester-indexed PLO features and removed 25 that were entirely missing, yielding 167 predictors; the 12 continuous targets were PLO1_Final to PLO12_Final. Table 2 reports summary statistics (n = 194): median attainment is highest for PLO8 (≈ 82.55) and PLO10 (≈ 82.38), while lower medians appear for PLO2 (≈ 70.43) and PLO1 (≈ 73.19). Dispersion varies: PLO7 shows the widest spread (SD ≈ 9.85), followed by PLO12 (SD ≈ 7.41), whereas PLO10 (SD ≈ 4.35) and PLO9 (SD ≈ 4.39) are comparatively tight. Several PLOs exhibit mild lower-tail skew (e.g., minima 39.15 for PLO7; 45.03 for PLO12), supporting the use of medians/interquartile ranges (IQRs) in descriptive summaries and MAE as the primary error metric on the 0-100 scale. Missingness was more noticeable in later semester features; zero-flags were restored to NaN before imputation, and all-NaN predictors were dropped. (For the 90/10 regime specifically, 13 constant features were removed post-imputation/scaling on the training fold, leaving 154 predictors for that split; this train-based decision was mirrored on the test set to avoid leakage.)

Table 2. Summary statistics for target variables in our dataset

	PLO 1	PLO 2	PLO 3	PLO 4	PLO 5	PLO 6	PLO 7	PLO 8	PLO 9	PLO 10	PLO 11	PLO 12
count	194	194	194	194	194	194	194	194	194	194	194	194
mean	73.18	70.67	76.52	74.02	77.15	80.64	77.90	81.71	76.72	81.96	81.19	78.16
std	6.77	7.01	5.98	5.06	4.63	4.79	9.85	6.44	4.39	4.35	5.30	7.41
min	52.31	50.52	59.08	60.27	59.33	59.90	39.15	63.06	61.43	59.17	64.11	45.03
25%	69.33	66.65	72.11	70.29	74.17	77.85	72.57	77.48	73.78	79.69	78.45	74.20
50%	73.19	70.43	77.66	74.30	77.60	81.34	78.16	82.55	76.84	82.38	81.71	79.52
75%	78.16	75.79	80.67	77.91	80.35	83.74	85.48	85.82	80.25	84.49	84.45	83.23
max	88.54	87.25	88.72	83.40	87.44	89.18	94.17	96.25	85.06	91.77	97.50	93.66

Three hold-out regimes were evaluated (70/30, 80/20, and 90/10). For each, we compared a panel of baseline learners using macro-averaged MAE, RMSE, and R^2 across the 12 PLO targets. Unless stated otherwise, all errors are reported in PLO score points on the original 0-100 scale (lower is better). Ten-fold cross-validation on the training portion was used to assess stability and support statistical comparisons; the fold-averaged CV results (mean ± SD) for each model and split are summarised in Table 3.

Table 3. Ten-fold cross-validation results (training folds) for final PLO prediction by model and hold-out regime (macro-averaged MAE↓, RMSE↓, R^2↑; mean ± SD, 0-100 scale).

70/30

	MAE Mean	MAE SD	RMSE Mean	RMSE SD	R^2 Mean	R^2 SD
CatBoost	2.287	0.314	3.342	0.644	0.663	0.069
Extra Trees	2.368	0.298	3.367	0.561	0.644	0.052
LightGBM	2.643	0.283	3.799	0.539	0.567	0.111
XGBoost	2.679	0.395	3.812	0.588	0.552	0.093
Gradient Boosting	2.339	0.306	3.378	0.559	0.645	0.073
HistGradient Boosting	2.571	0.287	3.736	0.523	0.583	0.093
Random Forest	2.596	0.333	3.680	0.655	0.583	0.063
AdaBoost	2.786	0.284	3.866	0.539	0.551	0.083
SVR	3.656	0.335	4.973	0.726	0.340	0.063
k-NN	3.249	0.383	4.423	0.689	0.415	0.076
Bayesian Ridge	**2.185**	0.365	**3.216**	0.734	**0.666**	0.092

80/20

	MAE Mean	MAE SD	RMSE Mean	RMSE SD	R^2 Mean	R^2 SD
CatBoost	2.158	0.260	3.242	0.478	0.700	0.051
Extra Trees	2.330	0.272	3.325	0.438	0.665	0.054
LightGBM	2.512	0.323	3.639	0.519	0.614	0.078
XGBoost	2.635	0.257	3.770	0.503	0.576	0.102
Gradient Boosting	2.242	0.215	3.286	0.401	0.684	0.056
HistGradient Boosting	2.441	0.277	3.564	0.480	0.638	0.068
Random Forest	2.540	0.254	3.618	0.420	0.612	0.054
AdaBoost	2.703	0.262	3.773	0.428	0.588	0.055
SVR	3.536	0.334	4.858	0.516	0.396	0.036
k-NN	3.213	0.404	4.416	0.596	0.445	0.095
Bayesian Ridge	**2.022**	0.286	**2.979**	0.508	**0.736**	0.063

90/10

	MAE Mean	MAE SD	RMSE Mean	RMSE SD	R^2 Mean	R^2 SD
CatBoost	2.071	0.303	3.054	0.535	0.717	0.074
Extra Trees	2.238	0.316	3.164	0.511	0.681	0.081
LightGBM	2.393	0.293	3.498	0.486	0.651	0.092
XGBoost	2.567	0.324	3.649	0.512	0.590	0.110
Gradient Boosting	2.149	0.274	3.105	0.527	0.707	0.072
HistGradient Boosting	2.339	0.304	3.424	0.510	0.663	0.089
Random Forest	2.454	0.318	3.469	0.524	0.631	0.081
AdaBoost	2.632	0.281	3.640	0.447	0.601	0.067
SVR	3.405	0.319	4.721	0.606	0.408	0.096
k-NN	3.098	0.403	4.276	0.590	0.457	0.125
Bayesian Ridge	**1.835**	0.249	**2.769**	0.512	**0.754**	0.097

70/30 Split

- Single models. Bayesian Ridge achieved the lowest hold-out error (MAE = 1.567, RMSE = 2.294, R^2 = 0.813), followed by CatBoost (MAE = 1.903, RMSE = 2.830, R^2 = 0.729), Extra Trees (MAE = 1.953, RMSE = 2.777, R^2 = 0.708), and Gradient Boosting (MAE = 1.974, RMSE = 2.836, R^2 = 0.713). SVR and k-NN had the largest errors.

- Ensembles. A convex blend of Bayesian Ridge (0.55) + CatBoost (0.35) + Gradient Boosting (0.10) delivered the best overall performance (MAE = 1.487, RMSE = 2.213, R^2 = 0.829), a ~5.1% MAE reduction vs. the best single model (Bayesian Ridge) and ~3.6% lower RMSE ($\Delta R^2 \approx +0.016$). A stacking variant did not surpass the single-model baseline (MAE ≈ 1.955).

- Cross-validation. Ten-fold CV on the training portion ranked Bayesian Ridge first (MAE = 2.185 ± 0.365), closely followed by CatBoost (2.287 ± 0.314), Gradient Boosting (2.339 ± 0.306), and Extra Trees (2.368 ± 0.298). An omnibus Friedman test rejected equal performance across models ($\chi^2 = 87.31$, $p = 1.83\times10^{-14}$), with top tree/boosting methods outperforming kernel/instance baselines, while the linear Bayesian Ridge emerged as the leading single learner under this data regime.

80/20 Split

- Single models. Bayesian Ridge gave the best hold-out performance (MAE = 1.458, RMSE = 2.110, R^2 = 0.798), followed by CatBoost (MAE = 1.668, RMSE = 2.404, R^2 = 0.768), Gradient Boosting (MAE = 1.702, RMSE = 2.364, R^2 = 0.751), and Extra Trees (MAE = 1.739, RMSE = 2.368, R^2 = 0.727). Tree/boosting methods remained competitive, while SVR and k-NN trailed with higher errors.

- Ensembles. A convex blend of Bayesian Ridge (0.55) + CatBoost (0.20) + Extra Trees (0.25) achieved the best overall result (MAE = 1.317, RMSE = 1.870, R^2 = 0.843), a ~9.7% MAE reduction relative to the best single model (Bayesian Ridge). A stacking variant did not surpass the single-model baseline (MAE ≈ 1.657).

- Cross-validation. Ten-fold CV on the training portion ranked Bayesian Ridge first (MAE = 2.022 ± 0.286), followed by CatBoost (2.158 ± 0.260) and Gradient Boosting (2.242 ± 0.215); Extra Trees was close behind (2.330 ± 0.272). A Friedman test rejected equal performance across models ($\chi^2 = 93.62$, $p = 1.02\times10^{-15}$), with tree/boosting methods consistently outperforming SVR/k-NN while the linear Bayesian Ridge remained a strong and here, leading baseline.

90/10 Split

- Single models. Bayesian Ridge achieved the lowest hold-out error (MAE = 1.363, RMSE = 1.900, R^2 = 0.726), followed by CatBoost (MAE = 1.429, RMSE = 1.947, R^2 = 0.764), Gradient Boosting (MAE = 1.492), and Extra Trees (MAE = 1.553).

- Ensembles. A convex blend of Bayesian Ridge (0.65) + CatBoost (0.35) + Extra Trees (0.00) delivered the best overall performance (MAE = 1.192, RMSE = 1.648, R^2 = 0.809), a ~12.6% MAE reduction versus the best single model. A simple stacking of the same trio did not surpass Bayesian Ridge (MAE = 1.444).

- Cross-validation. On 10-fold CV (training portion only), Bayesian Ridge ranked first by mean MAE (1.835 ± 0.249), followed by CatBoost (2.071 ± 0.303) and Gradient Boosting (2.149 ± 0.274). An omnibus Friedman test indicated significant differences across 11 learners ($\chi^2 = 96.8$, $p = 2.38\times10^{-16}$), with a similar pattern for RMSE and R^2.

Across all three hold-out regimes, the best single learner by MAE was Bayesian Ridge, with CatBoost a close runner-up (often yielding the highest R^2), followed by Gradient Boosting and Extra Trees. A simple convex blend of top models (Bayesian Ridge + CatBoost ± Extra Trees/Gradient Boosting) consistently improved hold-out MAE by ~5-13% over the best single learner (70/30: ~5.1%; 80/20: ~9.7%; 90/10: ~12.6%), whereas stacking with a linear meta-learner did not provide additional gains (Table 4). Ten-fold cross-validation on the training folds corroborated these patterns, placing Bayesian Ridge first by mean MAE across splits, with CatBoost close behind; Friedman tests rejected equal performance among learners in each regime ($\chi^2 \approx 87\text{-}97$, $p < 1\times10^{-14}$).

Table 4. Cross-split performance of single and ensemble models for final PLO prediction (MAE, RMSE, R²).

	70/30			**80/20**			**90/10**		
	MAE	RMSE	R2	MAE	RMSE	R2	MAE	RMSE	R2
Ensemble Convex Blend	**1.487**	2.213	**0.829**	**1.317**	1.870	**0.843**	**1.192**	1.648	**0.809**
Ensemble Stacking	1.955	2.761	0.708	1.657	2.223	0.744	1.444	1.892	0.727
CatBoost	1.903	2.830	0.729	1.668	2.404	0.768	1.429	1.947	**0.764**
Extra Trees	1.953	2.777	0.708	1.739	2.368	0.727	1.553	2.139	0.705
LightGBM	2.145	3.076	0.667	1.994	2.835	0.670	1.850	2.546	0.561
XGBoost	2.369	3.364	0.594	2.128	2.910	0.628	2.098	2.783	0.442
Gradient Boosting	1.974	2.836	0.713	1.702	2.364	0.751	1.492	2.048	0.714
HistGradient Boosting	2.121	2.999	0.681	1.957	2.790	0.680	1.719	2.357	0.622
Random Forest	2.166	3.031	0.670	1.944	2.649	0.699	1.751	2.361	0.637
AdaBoost	2.311	3.170	0.652	2.162	2.855	0.664	1.962	2.524	0.601
SVR	2.971	4.184	0.455	2.645	3.780	0.524	2.272	3.017	0.484
KNN	2.837	3.873	0.475	2.526	3.398	0.492	2.421	3.089	0.274
Bayesian Ridge	**1.567**	2.294	**0.813**	1.458	2.110	**0.798**	1.363	1.900	0.726

Feature importance analyses, using model native measures and SHAP values, provided insights into which semester-level learning outcomes drive final PLO scores.

Figures 1a-c report the Top-10 global SHAP importances for the Bayesian Ridge model under the 70/30, 80/20 and 90/10 splits. A consistent pattern emerges across splits: upper-year signals (Years 3-4) dominate, with Semester 2 features especially prominent. Seven predictors recur in the Top-10 for all three splits:

PLO7_Semester_2_Year_4, PLO1_Semester_2_Year_3, PLO2_Semester_1_Year_3,
PLO2_Semester_2_Year_3, PLO2_Semester_2_Year_1, PLO7_Semester_2_Year_2, and
PLO12_Semester_1_Year_4.

Figure 1. Champion model (Bayesian Ridge) global SHAP importances (Top-10) across hold-out regimes. Bars show within-split normalised mean |SHAP| for (a) 70/30, (b) 80/20, and (c) 90/10; upper-year (Years 3-4), Semester-2 signals are most prominent. See Table 5 for the cross-split consensus.

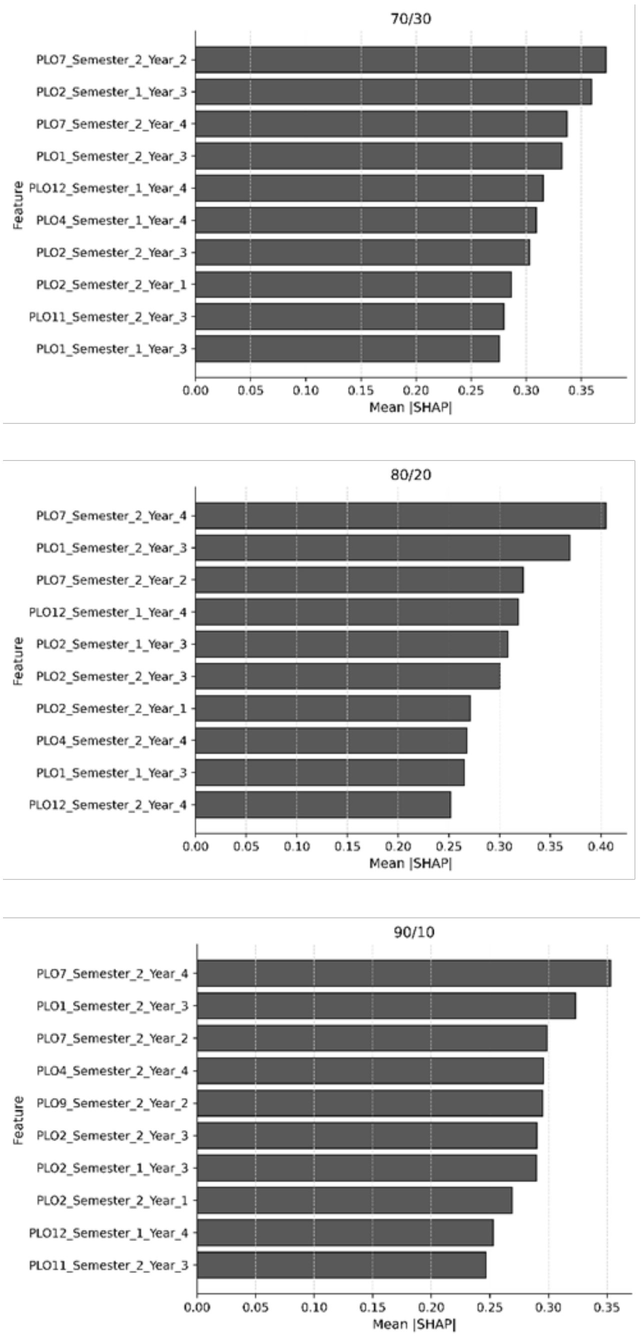

Table 5 shows a strong signal from the later years (third year and beyond) gradient in the champion's Top-10 SHAP features. Among the recurrent features present in all three splits (n = 7), Year 3 contributes 3, Year 4 contributes 2, and Years 1-2 contribute 1 each; Semester 2 accounts for 5 of the 7. The repeated presence of PLO2 (three signals) and PLO7 (two signals) suggests broad, transferable competencies that persist to graduation, while the inclusion of PLO2_Semester_2_Year_1 indicates a meaningful early-year signal. See Table 5 for the cross-split consensus.

Table 5. Cross-split consensus (Bayesian Ridge, Top-10): mean normalised SHAP importance and number of regimes (n = 3) in which each feature appears.

Rank	Feature	PLO	Year	Semester	Appears in Splits (max=3)	Mean Normalized Importance
1	PLO7_Semester_2_Year_4	7	4	2	3	0.02325
2	PLO1_Semester_2_Year_3	1	3	2	3	0.02173
3	PLO7_Semester_2_Year_2	7	2	2	3	0.02107
4	PLO2_Semester_1_Year_3	2	3	1	3	0.02028
5	PLO2_Semester_2_Year_3	2	3	2	3	0.01898
6	PLO12_Semester_1_Year_4	12	4	1	3	0.01877
7	PLO2_Semester_2_Year_1	2	1	2	3	0.01755
8	PLO4_Semester_2_Year_4	4	4	2	2	0.01216
9	PLO11_Semester_2_Year_3	11	3	2	2	0.01129
10	PLO1_Semester_1_Year_3	1	3	1	2	0.01119

Note. Features are ranked by mean normalised importance across the three hold-out regimes. "Appears in Splits" counts the number of regimes (max = 3) in which the feature is in the Top-10. Importance values are normalised (0-1) and rounded to five decimals.

Conclusions

This study demonstrates the feasibility of using multi-output regression to predict continuous PLO attainment scores. Findings are based on 194 anonymised students and 167 semester-indexed predictors spanning each student's full enrolment (typically eight semesters, sometimes more). We then evaluated a broad suite of machine learning algorithms: CatBoost, Extra Trees, Gradient Boosting, LightGBM, Random Forest, XGBoost, k-nearest neighbours, support vector regression, and Bayesian Ridge, under multiple train-test splits with cross-validation.

Predictive performance (RQ1-RQ2). On the 0-100 PLO scale, multi-output regression achieved strong accuracy across all three regimes, with Bayesian Ridge consistently delivering the lowest MAE: 1.567 (70/30), 1.458 (80/20), and 1.363 (90/10). A simple convex blend of top learners (Bayesian Ridge + CatBoost ± Extra Trees/Gradient Boosting) improved hold-out MAE by ~5-13% over the best single model (70/30: ~5.1%; 80/20: ~9.7%; 90/10: ~12.6%), while stacking did not add further benefit. By R^2, CatBoost occasionally led (notably in 90/10), but Bayesian Ridge provided the most reliable MAE overall. Ten-fold CV on the training folds corroborated these rankings (Friedman tests, $p < 1\times10^{-14}$).

Drivers and implications (RQ3). SHAP analyses of the champion indicate a strong signal from the later years (third year and beyond), with Semester-2 features most prominent and recurrent contributions from PLO2 and PLO7. Importantly, an early-year indicator (e.g., PLO2_Semester_2_Year_1) remains informative, suggesting that earlier interventions on foundational outcomes can improve end-of-programme attainment. Jointly modelling all twelve PLOs leverages inter-outcome correlations and aligns prediction with accreditation-relevant competencies. The data cleaning protocol (zero→NaN restoration, out-of-range capping, Year ≥ 3 completion screen, fold-confined imputation/scaling) further improved robustness and interpretability.

A substantial share of the predictor matrix reflects structural, curriculum-driven missingness (e.g., course-PLO mappings and assessment offerings vary by semester, cohort, and programme), so many PLO-semester cells were never observed for particular students. We used zeros only as temporary placeholders for reshaping, then restored them to NaN before modelling; out-of-range PLO values were capped to 100 to preserve the 0-100 scale. Missing predictors were handled via fold-confined median imputation with MinMax scaling, and we dropped all-NaN features, removed constant features on the training fold, and excluded rows with incomplete target, steps that reduce leakage but cannot eliminate potential bias if the missingness is Missing Not at Random (MNAR) or MNAR-like. External validity is constrained by the single-institution, one programme, 2015-2019, local-only cohort, and by small test sets (e.g., 90/10 has n=20), which can inflate split-specific variability despite the use of 10-fold CV. Finally, fairness was not assessed; if demographic or socio-economic disparities exist, algorithmic bias could remain unobserved and should be evaluated before any deployment or policy use.

Future work should expand both the data set and the feature space. Adding later cohorts (e.g., 2020+), other departments, and ideally multi-institution data will improve sample size, diversity, and external validity. Enriching predictors with demographic and socio-economic attributes, behavioural traces (LMS interaction, attendance, assessment timing), and psychological/affective indicators (e.g., grit, self-efficacy)-alongside curriculum metadata (course-PLO mappings, assessment modality), can raise accuracy and enable formal fairness analyses. Methodologically, we recommend time-aware and programme held-out validation to test generalisation, multiple imputation plus sensitivity analyses for MNAR-like missingness, calibration and uncertainty quantification (e.g., conformal prediction), and fairness auditing with mitigation (reweighting or group-conditional calibration). These extensions would support earlier, better-targeted interventions while ensuring results remain robust, transparent, and equitable.

References

Ab Rahim, A. A., & Buniyamin, N. (2022). Predicting Engineering Students' Academic Performance using Ensemble Classifiers - A Preliminary Finding. Journal of Electrical & Electronic Systems Research, 20(APR2022), 92–101. https://doi.org/10.24191/jeesr.v20i1.013

Abiodun, O. J., & Wreford, A. I. (2024). Student's Performance Evaluation Using Ensemble Machine Learning Algorithms. Engineering and Technology Journal, 09(08). https://doi.org/10.47191/etj/v9i08.23

Agha, D., Meghji, A. F., Bhatti, S., & Memon, M. (2023). Educational Data Mining in Outcome-Based Education: An Analysis of Predictive Models for Program Learning Outcome Attainment. VAWKUM Transactions on Computer Sciences, 11(2), 123–138. https://doi.org/10.21015/vtcs.v11i2.1706

Almaghaslah, D. (2025). Challenging the curve: can ChatGPT-generated MCQs reduce grade inflation in pharmacy education. Frontiers in Pharmacology, 16(January), 1–7. https://doi.org/10.3389/fphar.2025.1516381

Alturki, S., Cohausz, L., & Stuckenschmidt, H. (2022). Predicting Master's students' academic performance: an empirical study in Germany. Smart Learning Environments, 9(1). https://doi.org/10.1186/s40561-022-00220-y

Asiah, M., Nik Zulkarnaen, K., Safaai, D., Nik Nurul Hafzan, M. Y., Mohd Saberi, M., & Siti Syuhaida, S. (2019). A Review on Predictive Modeling Technique for Student Academic Performance Monitoring. MATEC Web of Conferences, 255, 03004. https://doi.org/10.1051/matecconf/201925503004

Czerkawski, B. C. (2015). When Learning Analytics Meets E-Learning. Online Journal of Distance Learning Administration, 18(2), 1–5. http://tinyurl.com/acz5f7o

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. Frontiers of Computer Science, 14(2), 241–258. https://doi.org/10.1007/s11704-019-8208-z

E. Ben George, Senthilkumar, R., Al-Junaibi, F., & Al-Shuaibi, Z. (2025). Explainable AI Methods for Predicting Student Grades and Improving Academic Success. Journal of Information Systems Engineering and Management, 10(23s), 117–126. https://doi.org/10.52783/jisem.v10i23s.3680

El Jihaoui, M., Abra, O. E. K., & Mansouri, K. (2025). Predicting and Interpreting Student Academic Performance: A Deep Learning and Shapley Additive Explanations Approach. SHS Web of Conferences, 214, 01001. https://doi.org/10.1051/shsconf/202521401001

Falát, L., & Piscová, T. (2022). Predicting GPA of University Students with Supervised Regression Machine Learning Models. Applied Sciences (Switzerland), 12(17). https://doi.org/10.3390/app12178403

Ghimire, S., Abdulla, S., Joseph, L. P., Prasad, S., Murphy, A., Devi, A., Barua, P. D., Deo, R. C., Acharya, R., & Yaseen, Z. M. (2024). Explainable artificial intelligence-machine learning models to estimate overall scores in tertiary preparatory general science course. Computers and Education: Artificial Intelligence, 7(November), 100331. https://doi.org/10.1016/j.caeai.2024.100331

Jaehn, M., Hissbach, J., Frickhoeffer, M., Weppert, D., Zimmerhofer, A., Hampe, W., Kadmon, M., & Becker, N. (2025). Predictive validity of admission tests and educational attainment on preclinical academic performance

– a multisite study. BMC Medical Education, 25(1). https://doi.org/10.1186/s12909-025-07974-2

Jović, J., Kisić, E., Milić, M. R., Domazet, D., & Chandra, K. (2022). Prediction of student academic performance using machine learning algorithms. CEUR Workshop Proceedings, 3454(September), 31–39.

Kamal, P., & Ahuja, S. (2019). Academic performance prediction using data mining techniques: Identification of influential factors effecting the academic performance in undergrad professional course. In Advances in Intelligent Systems and Computing (Vol. 741). Springer Singapore. https://doi.org/10.1007/978-981-13-0761-4_79

Khan, H. F., Qayyum, S., Beenish, H., Khan, R. A., Iltaf, S., & Faysal, L. R. (2025). Determining the alignment of assessment items with curriculum goals through document analysis by addressing identified item flaws. BMC Medical Education, 25(1). https://doi.org/10.1186/s12909-025-06736-4

Mio, M. J. (2024). Alternative grading strategies in organic chemistry: a journey. Frontiers in Education, 9(May), 1–13. https://doi.org/10.3389/feduc.2024.1400058

Ngulub, P., & Masumbika Ncube, M. (2025). Predicting Academic Success and Identifying At-Risk Students: A Systematic Review of Data Analytics and Machine Learning Approaches in Higher Education Institutions. Educational Administration: Theory and Practice, January. https://doi.org/10.53555/kuey.v31i1.8447

Nor Afiqah Wan Othman, W., Abdullah, A., & Romli, A. (2020). Predicting Graduate Employability based on Program Learning Outcomes. IOP Conference Series: Materials Science and Engineering, 769(1). https://doi.org/10.1088/1757-899X/769/1/012018

Olaleye, T. O., & Vincent, O. R. (2020). A Predictive Model for Students’ Performance and Risk Level Indicators Using Machine Learning. 2020 International Conference in Mathematics, Computer Engineering and Computer Science, ICMCECS 2020. https://doi.org/10.1109/ICMCECS47690.2020.240897

Pali, P., & Verma, S. (2024). Predictive Analytics for Student Performance: A Machine Learning Model for Higher Education. International Journal of Innovative Research in Computer and Communication Engineering, 12(05), 8151–8158. https://doi.org/10.15680/ijircce.2024.1205366

Pao Liew, C., & Loo Kiew, P. (2022). Sustainable Assessment: The Inevitable Future of Engineering Curriculum. ASEAN Journal of Engineering Education, 6(1), 23–32.

Park, B., & Cho, J. (2023). How does grade inflation affect student evaluation of teaching? Assessment and Evaluation in Higher Education, 48(5), 723–735. https://doi.org/10.1080/02602938.2022.2126429

Rafiq, J. E., Abdelali, Z., Amraouy, M., Nouh, S., & Bennane, A. (2025). Predicting academic performance: toward a model based on machine learning and learner’s intelligences. International Journal of Electrical and Computer Engineering, 15(1), 645–653. https://doi.org/10.11591/ijece.v15i1.pp645-653

Raji, N. R., Kumar, R. M. S., & Biji, C. L. (2024). Explainable Machine Learning Prediction for the Academic Performance of Deaf Scholars. IEEE Access, 12(February), 23595–23612. https://doi.org/10.1109/ACCESS.2024.3363634

Rane, N., Choudhary, S. P., & Rane, J. (2024). Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions. Studies in Medical and Health Sciences, 1(2), 18–41. https://doi.org/10.48185/smhs.v1i2.1225

Sghir, N., Adadi, A., & Lahmer, M. (2023). Recent Advances in Predictive Learning Analytics: A Decade Systematic Review (2012–2022). In Education and Information Technologies (Vol. 28, Issue 7). Springer US. https://doi.org/10.1007/s10639-022-11536-0

Tjandra, E., Ferdiana, R., & Setiawan, N. A. (2024). OBE-Based Course Outcomes Prediction Using Machine Learning Algorithms. 2024 International Conference on Intelligent Cybernetics Technology and Applications, ICICyTA 2024, March, 197–202. https://doi.org/10.1109/ICICYTA64807.2024.10913307

Wang, C., Chen, Q., Xue, B., & Zhang, M. (2024). Multi-task Genetic Programming with Semantic based Crossover for Multi-output Regression. GECCO 2024 Companion - Proceedings of the 2024 Genetic and Evolutionary Computation Conference Companion, 543–546. https://doi.org/10.1145/3638530.3654282

Xue, H., & Niu, Y. (2023). Multi-Output Based Hybrid Integrated Models for Student Performance Prediction. Applied Sciences (Switzerland), 13(9). https://doi.org/10.3390/app13095384

Yekun, E. A., & Haile, A. T. (2021). Student Performance Prediction with Optimum Multilabel Ensemble Model. Journal of Intelligent Systems, 30(1), 511–523. https://doi.org/10.1515/jisys-2021-0016

Zanellati, A., Zingaro, S. P., & Gabbrielli, M. (2024). Balancing Performance and Explainability in Academic Dropout Prediction. IEEE Transactions on Learning Technologies, 17, 2140–2153. https://doi.org/10.1109/TLT.2024.3425959

Zayani, H. M., Abdelfattah, W., Slimane, J. Ben, Kachoukh, A., & Sellami, R. (2024). A Framework for Efficient and Accurate Automated CLO and PLO Assessment. Engineering, Technology and Applied Science Research, 14(2), 13362–13368. https://doi.org/10.48084/etasr.6846

Zhang, R., Zhou, P., & Chai, T. (2023). Improved Copula-based conformal prediction for uncertainty quantification of multi-output regression. Journal of Process Control, 129, 103036. https://doi.org/10.1016/j.jprocont.2023.103036

Zhu, W. (2024). High school student GPA prediction by various linear regression models. Theoretical and Natural Science, 52(1), 153–162. https://doi.org/10.54254/2753-8818/52/2024ch0136