Journal of
Data Science & Applied Analytics

Copyright © Authors, 2026

Article

Handling Imbalanced Ordinal Grade Prediction in Educational Analytics Using Weighted Ordinal Logistic Regression

Dipesh Baral[1]*, Dindu Lama[2], Dinesh Baral[3], and Priyanshu Kharel[4]

1 Department of Mathematics and Statistics, Washington State University, USA
2 Department of Business and Economics, Wittenberg University, USA
3 Independent Researcher, USA
4 Institute of Science and Technology, Tribhuvan University, Nepal
*Correspondence: Dipesh.baral@wsu.edu

Abstract

Imbalanced ordinal outcomes are common in educational analytics, where lower academic grades occur less frequently yet remain critical for identifying at-risk students. Standard predictive models often favor majority outcomes, reducing sensitivity to minority grade categories. This study evaluates weighted ordinal logistic regression as an interpretable approach for handling imbalanced ordinal grade prediction using student gender, age, and math anxiety as predictors. Unweighted and inverse-frequency weighted models were compared using precision, recall, F1 score, accuracy, mean absolute error (MAE), quadratic weighted kappa (QWK), and area under the curve (AUC). Weighting improved detection of minority lower-grade categories (C–E), increasing macro F1, but reduced performance for majority grades and overall discriminative ability. These findings suggest that weighted ordinal logistic regression can support educational risk identification when minority-class detection is prioritized, although predictive performance remains constrained by limited sample size and predictor strength.

Keywords: Imbalanced ordinal classification; Educational data mining; Grade inflation; Ordinal logistic regression; Predictive modeling

1. Introduction

Grades are widely recognized as indicators of academic achievement, progression, and institutional decision-making. Over the past several decades, however, grade distributions in higher education have shifted away from the traditional bell-shaped pattern toward concentrations of higher grades, with A and B grades increasingly dominating while lower grades become comparatively rare (Arthurs et al., 2019; Rojstaczer & Healy, 2012). Evidence from multiple institutions has documented substantial increases in top grades alongside declines in lower grades (Achen & Courant, 2009; Jeong et al., 2024; Levine & Cureton, 1998), a pattern further intensified during and after the COVID-19 period (Rodríguez-Planas, 2022; Ciftci & Karadağ, 2024). The persistence of this issue remains evident, as Harvard University recently approved a grading reform that will cap A grades in undergraduate courses beginning in 2027, reflecting renewed institutional concern over grade inflation and grade compression

Received: date
Revised: date
Accepted: date

Open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license.

(Gibson, 2026). Although grading systems differ across institutions and rising grades do not automatically indicate artificial inflation (Jephcote et al., 2021), the broader empirical pattern suggests that higher grades have become increasingly prevalent, reducing differentiation among lower-performing students (Pattison et al., 2013; Tucker & Courts, 2010; Jewell et al., 2013; Pearce, 2017; Kamber, 2008).

Despite their increasing rarity, lower grades remain critically important in higher education. They serve as gatekeepers for academic progression, influencing whether students can advance through prerequisite coursework, remain in competitive programs, or qualify for academic opportunities (Husong, 2024; Jenkins et al., 2009). Low grades also function as indicators for academic intervention, often triggering referrals to developmental coursework, support services, or remediation programs (Sanabria et al., 2020; Ran & Lin, 2022; Kearney et al., 2024). Beyond progression and remediation, poor academic performance remains a strong predictor of student attrition, particularly when academic difficulties emerge early in the educational trajectory (Gershenfeld et al., 2016; Martinez et al., 2009). Consequently, identifying students at risk of lower academic outcomes remains an important problem in educational analytics.

Predictive modeling has increasingly been used to identify academically at-risk students (Almalawi et al., 2024; Polyzou & Karypis, 2019). Machine learning approaches, including decision trees, support vector machines, and random forests, have been applied in this context (Wongvorachan & Bulut, 2023; Bujan et al., 2022; Abdulazeez & Abdulwahab, 2018; Limanto & Saikhu, 2024; Albrieki et al., 2021; Saa, 2016). However, predictive modeling in educational settings frequently encounters class imbalance, where high grades dominate historical data while lower-grade outcomes remain underrepresented. Common responses include resampling and synthetic oversampling techniques such as SMOTE (Chawla, 2002), but such methods may introduce artificial observations while failing to preserve the ordinal structure inherent in grade outcomes. Furthermore, many predictive models offer limited interpretability, reducing their practical utility for educational stakeholders seeking transparent and actionable decision-support tools.

Ordinal logistic regression (OLR) provides an interpretable modeling approach for ordered categorical outcomes, preserving the ordinal structure of grades rather than treating them as nominal categories (Adejumo, 2013; Baral et al., 2025; Bista et al., 2025). Weighting strategies have also been incorporated into ordinal models to address class imbalance by increasing the influence of underrepresented outcome categories (Zhou et al., 2008; Saputro & Widyaningsih, 2017; Abreu et al., 2008). Previous studies have applied such models using demographic and academic predictors, including socioeconomic status, prior academic performance, and gender (Sesay et al., 2021; Muya, 2018; Carophine, 2016; Dadzie et al., 2025). However, psychological predictors such as math anxiety remain comparatively underutilized despite strong evidence linking them to academic performance and persistence in mathematics-related disciplines (Barroso et al., 2021; Baral, 2025; Namkung et al., 2019).

This study evaluates inverse-frequency weighted ordinal logistic regression as an interpretable predictive modeling approach for imbalanced ordinal grade prediction. Using student age, gender, and math anxiety as predictors, the study compares weighted and unweighted ordinal logistic regression in their ability to identify lower-grade outcomes. Building on prior work applying unweighted ordinal logistic regression to math anxiety and academic performance (Baral et al., 2025), the present study extends that work by explicitly addressing class imbalance within an educational predictive analytics context.

2. Research Methods

Data and Methodology

The dataset was obtained from the corresponding author of Baral et al., (2025). It consists of undergraduate students' demographic and psychological predictors (e.g., gender, age, anxiety) and their academic grades, coded as an ordinal outcome variable with five ordered categories (E, D, C, B, A). Details about the data and data collection process have been discussed in Baral et al., (2025). Here, we emphasize that the distribution of grades reveals a pronounced skew toward higher grades, with A (16) and B (14) accounting for 62.5% of the sample, while lower grades, E (6) and D (4), constitute 20.8%, and C (8) represents 16.7%, making it appropriate for our study.

An ordinal logistic regression (OLR) model was estimated under the proportional odds assumption, with academic grade as the ordinal outcome variable and anxiety (continuous), age (categorical), and gender (categorical) entered as predictors. The cumulative logit model is expressed as:

$$\log\frac{Pr(grade > k)}{Pr(grade \leq k)} = \theta_k - (\beta_1 \times gender + \beta_2 \times age + \beta_3 \times anxiety)$$

where θ_k (with k =1, 2, 3, and 4) are the intercepts for each grade level boundary (A/B, B/C, C/D, and D/E) and β_1, β_2 and β_3 are the coefficients for gender, age, and math anxiety, respectively. The model predicts the log-odds of being in a category greater than *k*.

To improve the detection of minority grade levels, weights are incorporated into the likelihood function, adjusting the contribution of each observation based on its category frequency. For total observations, let $(\text{freq}(Grade_k))$ represents the observed frequency of the grade category *k*. Then, the weights for each observation of category *k* are defined as under:

(i) Unweighted scheme with $w_k = 1$ such that all observations are treated equally.

(ii) Full inverse weighted with $w_k = \frac{N}{\text{freq}(Grade_k)}$ such that the assigned weight is inversely proportional to the category frequency. This approach amplifies the contribution of underrepresented categories while attenuating the influence of prevalent ones. For instance, for the category D with frequency 4, the weight $w_D = 48/4 = 12$ whereas for the category A with frequency 16, the weight $w_A = 48/16 = 3.$

The weighting scheme modifies the corresponding likelihood function from

$$L = \prod_{i=1}^{N} P(Grade_i = k_i \mid \boldsymbol{x_i}) \text{ to } L_w = \prod_{i=1}^{N} P(Grade_i = k_i \mid \boldsymbol{x_i})^{w_i}$$

thereby enhancing the contribution of the rare grade level in the likelihood space. Here, *i* ranges from 1 to 48 for the 48 observations.

Model performance was evaluated using both conventional classification metrics (precision, recall, F1 score, and accuracy) and ordinal-specific metrics (mean absolute error (MAE) on ranks and quadratic weighted kappa (QWK)). In addition, receiver operating characteristic area under the curve (ROC–AUC) was estimated out-of-fold (OOF) using stratified K-fold cross-validation to provide robust measures of discriminative capacity. A brief description of the metrics is provided below.

1) **Accuracy:** Overall accuracy measures the proportion of correctly classified observations across all classes and is defined as

$$Accuracy = \frac{\sum_{k=1}^{K} TP_k}{N},$$

Where TP_k denotes true positives for class *k* and *N* is total observations.

2) **Precision:** Precision for class *k* is defined as

$$Precision_k = \frac{TP_k}{TP_k + FP_k},$$

where TP_k is the number of true positives for the class *k* and FP_k the false positives. High precision indicates that when the model predicts class *k*, it is usually correct (few false alarms).

3) **Recall:** Recall for class *k* is defined as

$$Recall_k = \frac{TP_k}{(TP_k + FN_k)},$$

where FN_k denotes the false negatives. High recall indicates that most actual instances of class *k* are successfully identified (few misses).

4) **F1 score:** F1 score for class *k* is the harmonic mean of precision and recall of class *k* and is defined as

$$F1_k = \frac{2(Precision_k * Recall_k)}{(Precision_k + Recall_k)}.$$

A high F1 indicates a good balance between catching true cases and avoiding false alarms.

5) **Macro vs. Micro Averaging.** In multiclass settings, performance can be summarized using either macro- or micro-averaging. Macro-averaging computes metrics independently for each class and reports their unweighted mean, giving equal weight to both frequent and rare classes. Micro-averaging pools all true positives, false positives, and false negatives before computing the metric. In other words, macro scores highlight balance across classes, whereas micro scores emphasize aggregate correctness dominated by majority classes.

6) **Mean Absolute Error** (**MAE):** MAE on ranks measures the average absolute difference between predicted and observed grade positions and is defined as

$$\text{MAE} = \left(\frac{1}{N}\right) \sum_{i=1}^{N} |\hat{y}_i - y_i|,$$

where $\hat{y}_i$ and y_i are integer-encoded predicted and observed grades. A low MAE indicates that predictions are close in rank to the true grade.

7) **Quadratic Weighted Kappa** (**QWK):** Quadratic Weighted Kappa compares observed agreement to expected agreement, with heavier penalties for larger ordinal disagreements, and is defined as

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{ij}\, O_{ij}}{\sum_{i,j} w_{ij}\, E_{ij}}, \quad with \quad w_{ij} = \frac{(i-j)^2}{(K-1)^2}.$$

Here, O is the observed confusion matrix, E is the expected matrix under chance, and K is the number of classes. High QWK indicates strong agreement between predicted and actual grades, considering the order of the grades.

8) Receiver Operating Characteristic (ROC) and Area Under the Curve (**AUC):**

ROC curves were constructed in a one-vs-rest manner for each grade. The corresponding per-class AUC quantifies the model's ability to discriminate a given grade from all others. A macro-AUC was computed as the unweighted mean of per-class AUCs (treating all grades equally), while a micro-AUC was computed by pooling predictions across all classes to provide a global measure of discriminative capacity.

3. Results

Weighted and unweighted ordinal logistic regression (cumulative link model) were compared for imbalanced ordinal grade prediction using gender, age group, and math anxiety as predictors.

Classification performance is summarized in Figure 1 and Table 2. The unweighted model showed limited ability to discriminate across grade categories, with predictive performance concentrated primarily in the majority classes. Its macro F1 score was 0.20, indicating poor balance across classes, while the micro F1 score (approximately equivalent to overall accuracy in this multiclass setting) reached 0.33. In contrast, the inverse-frequency weighted model improved recognition across underrepresented grade categories, increasing macro F1 to 0.36 and modestly improving micro F1 to 0.38. Quadratic weighted kappa (QWK) also increased slightly from 0.19 to 0.25, indicating modestly better ordinal agreement between predicted and observed grades. However, this improvement came at the cost of increased ordinal prediction error, with mean absolute error (MAE) rising from 1.17 to 1.33.

Table 1

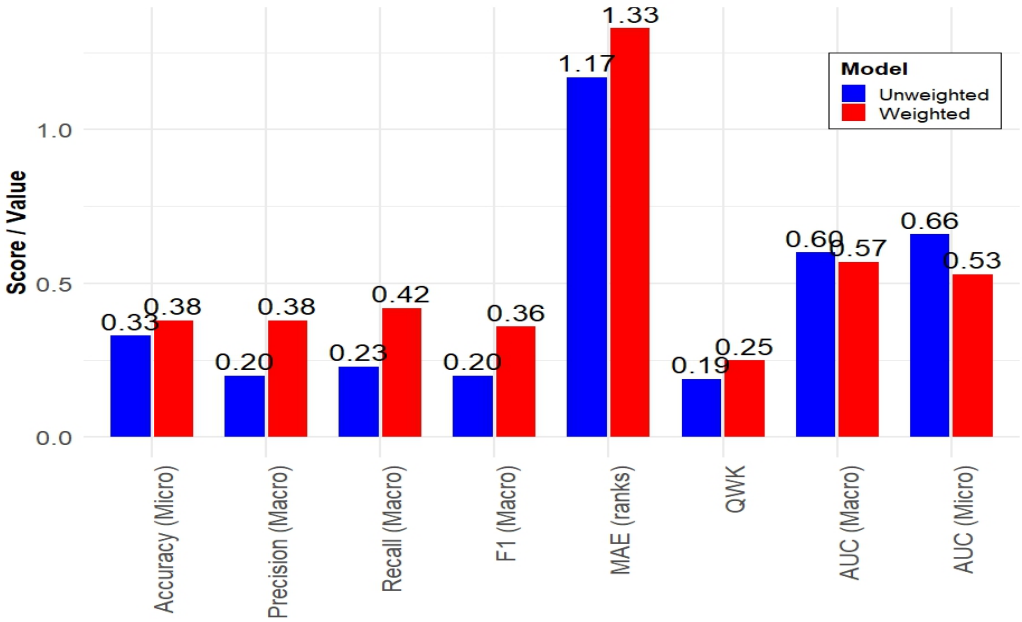
Model fit and parameters of the unweighted and weighted ordinal logistic regression.

Coefficients	Estimate		SE		Wald test p-value		$\chi2$ p-value	
	UW	W	UW	W	UW	W	UW	W
Gender (Female)	-0.53	-0.80	0.60	0.27	0.373	0.668	0.371	0.000
Age (18-25)	-0.41	-1.67	0.90	0.39	0.652	0.537	0.651	0.000
Anxiety	-0.03	-0.04	0.01	0.01	0.097	0.439	0.092	0.000
Intercept E|D	-4.58	-6.11	1.78	5.60	0.010	0.276	NA	NA
Intercept D|C	-3.92	-4.94	1.73	5.28	0.023	0.350	NA	NA
Intercept C|B	-3.03	-4.01	1.69	5.12	0.073	0.434	NA	NA
Intercept B|A	-1.75	-2.97	1.68	5.08	0.297	0.560	NA	NA

Note. Model fit and parameter outputs for the unweighted and weighted ordinal logistic regression. UW indicates unweighted model statistics, W indicates weighted model statistics and $\chi2$ p-value is the drop-1 $\chi2$ p-value. The Likelihood Ratio (LR) $\chi2$ vs intercept only was 4.947 (df=3, p=0.176) for the unweighted model, while it was 127.974 (df=3, p=0.000) for the weighted model. Similarly, the McFadden R^2, Cox-Snell R^2, and Nagelkerke R^2 were 0.0346, 0.0979, and 0.1032, respectively, for the unweighted model, and those for the weighted model were 0.8939, 0.9305, and 0.9801, respectively.

Figure 1

Overall classification metrics for the unweighted and weighted models



Note. Bar plot of Precision, Recall, F1, MAE, QWK, and AUC values obtained for each class of the weighted model and the unweighted model. UW stands for unweighted, and W stands for weighted.

Table 2

Per-class classification metrics for the unweighted and weighted models

Class	Precision		Recall		F1		AUC	
	UW	W	UW	W	UW	W	UW	W
E	0.33	0.36	0.17	0.67	0.22	0.47	0.70	0.78
D	0.00	0.18	0.00	0.50	0.00	0.27	0.54	0.52
C	0.00	0.60	0.00	0.38	0.00	0.46	0.59	0.53
B	0.20	0.33	0.29	0.07	0.24	0.12	0.61	0.51
A	0.44	0.44	0.69	0.50	0.54	0.47	0.52	0.53

Note. Precision, Recall, F1, and AUC values obtained for each class of the weighted model and the unweighted model. UW stands for unweighted, and W stands for weighted.

Table 2 presents class-specific predictive performance. The weighted model substantially improved detection of minority grades, particularly grade E, where F1

increased from 0.22 to 0.47 and AUC from 0.70 to 0.78. Minority grades C and D, which were not identified by the unweighted model, achieved F1 scores of 0.46 and 0.27, respectively, under inverse weighting. These findings suggest that weighting improved sensitivity to rare grade outcomes that were otherwise largely overlooked.

This improvement, however, involved clear trade-offs. Predictive performance for majority grades declined under weighting, with the F1 score for grade A decreasing from 0.54 to 0.47 and grade B declining from 0.24 to 0.12. Similarly, while the weighted model improved discrimination for the rarest category, aggregate discriminative performance weakened, as reflected by the decline in micro AUC from 0.66 to 0.53. Taken together, these findings indicate that inverse-frequency weighting redistributed predictive emphasis toward minority classes while reducing overall ranking performance.

For interpretability, model coefficients and fit statistics are presented in Table 1. The unweighted model showed limited improvement over the intercept-only model, $\chi^2(3) = 4.95$, $p = 0.176$, with low explanatory power (Nagelkerke $R^2 = 0.10$). In contrast, the weighted model showed substantially improved fit according to weighted likelihood criteria, $\chi^2(3) = 127.97$, $p < .001$, with markedly higher pseudo-R^2 values. Drop-1 likelihood ratio tests suggested that gender, age, and anxiety contributed to the weighted model. Given the predictive focus of the present study, these coefficient estimates primarily provide interpretive context rather than serving as the principal basis for model comparison.

4. Discussion

This study evaluated whether inverse-frequency weighting improves predictive performance for minority classes in an imbalanced ordinal classification setting. The findings indicate that weighting substantially improved the detection of underrepresented grade categories (E, D, and C), which were poorly recognized under theunweighted model. In particular, grade E showed meaningful gains in both F1 score and AUC, suggesting improved discrimination for the rarest outcome category. These results indicate that inverse-frequency weighting can effectively redistribute predictive emphasis toward minority ordinal outcomes in imbalanced classification problems.

The trade-off, however, was equally evident. While minority-class recognition improved, predictive performance for majority grades declined, and aggregate discrimination metrics—particularly micro AUC—moved closer to chance levels. This pattern reflects a familiar tension in imbalance-aware predictive modeling: improving sensitivity to rare but consequential outcomes often comes at the expense of global ranking performance and overall discriminative reliability. Whether this trade-off is acceptable depends on the intended application. In educational early-risk detection, prioritizing recall and F1 for minority low-grade outcomes may be justified because failing to identify struggling students may delay intervention or academic support. In contrast, in settings where accurate differentiation among majority or high-performing categories is equally important—such as honors classification, scholarship decisions, or rank-sensitive evaluations—full inverse weighting may distort predictive priorities and reduce overall utility.

From an applied analytics perspective, the findings also underscore the importance of evaluating imbalanced ordinal predictive models using multiple complementary performance metrics. Class-specific precision, recall, and F1 scores provide insight into minority-class detection, whereas aggregate discrimination metrics such as macro and micro AUC offer a broader threshold-independent assessment of model behavior. Reliance on a single metric may obscure important trade-offs between sensitivity to rare outcomes and overall predictive discrimination. Transparent reporting of both class-level

and aggregate performance measures is therefore essential for evaluating imbalance-aware predictive models intended for decision-support applications.

Two methodological choices strengthen confidence in the present findings. First, performance was evaluated using both threshold-dependent metrics (precision, recall, F1, macro and micro averages) and threshold-independent discrimination metrics (per-class and aggregate AUC), allowing performance trade-offs to be examined from complementary perspectives. Second, AUC estimates were derived from out-of-fold predictions under stratified cross-validation, reducing the likelihood that observed performance reflects simple in-sample overfitting. Together, these design choices provide a more balanced evaluation of weighting effects than reliance on a single metric.

At the same time, overall predictive performance remained limited. QWK values remained below 0.30, most class-level AUC values were near chance, and MAE exceeded one grade level, indicating that the available predictor set provided only modest predictive signal. These findings suggest that weighting alone cannot compensate for weak or sparse feature information; rather, it changes how predictive emphasis is distributed across outcome classes. More informative feature spaces—including prior academic performance, attendance, course enrollment history, instructional context, and additional psychological or motivational measures—may be necessary to achieve practically useful ordinal grade prediction.

Although this study is situated in an educational analytics context, the underlying methodological challenge extends beyond education. Similar imbalance-related trade-offs arise in healthcare risk stratification, severity classification, customer satisfaction scoring, and other applied predictive settings involving rare but consequential ordinal outcomes. In such contexts, interpretable weighting-based approaches may offer a transparent alternative when sensitivity to minority outcomes is prioritized over aggregate discrimination performance. should discuss the results and how they can be interpreted from the perspective of previous studies and of the working hypotheses.

5. Limitations and Extensions

Several limitations should be acknowledged. First, the relatively small sample size likely increased variance and contributed to unstable per-class performance estimates, particularly for the rarest outcome categories. Although stratified cross-validation was used to reduce overfitting risk in discrimination estimates, external validation using larger and more diverse datasets will be necessary to assess generalizability and robustness. Bootstrap resampling may also provide useful supplementary uncertainty estimates in future analyses.

Second, attendance and withdrawal status were unavailable in the present dataset. As a result, some lower-grade outcomes may reflect administrative non-completion rather than genuine academic underperformance. Future work should incorporate attendance and withdrawal indicators to better distinguish disengagement-related outcomes from academic performance itself.

Third, this study evaluated full inverse-frequency weighting as a transparent baseline strategy for handling ordinal class imbalance. While effective for improving minority-class detection, this approach produced notable trade-offs, including reduced performance in majority categories and weaker global discrimination. Alternative imbalance-aware strategies—including moderated weighting schemes, cost-sensitive loss formulations, focal loss adaptations, and ordinal resampling approaches—may yield more balanced performance by avoiding the extreme reweighting imposed by full inverse-frequency weighting.

Fourth, calibration performance was not explicitly assessed. Although discrimination metrics such as AUC and F1 are useful for evaluating classification behavior, they do not

indicate whether predicted probabilities align well with observed outcome frequencies. Future work should incorporate calibration diagnostics such as calibration curves, Brier scores, and expected calibration error (ECE), particularly if predicted probabilities are intended for intervention prioritization or other downstream decision-support applications.

Several methodological extensions also warrant consideration. Weighting strength may be treated as a tunable hyperparameter rather than fixed at full inverse-frequency weighting, for example through schemes of the form $w \propto 1/freq^{\alpha}$, where α is optimized to balance minority-class sensitivity against overall discrimination. Additional ordinal-aware evaluation metrics, including concordance indices or Somers' D, may complement conventional classification measures by more effectively capturing ordered prediction quality across adjacent categories.

Finally, although the present application focuses on educational analytics, similar imbalance challenges arise in other applied domains involving ordinal risk prediction. Comparable predictive trade-offs are relevant in healthcare risk stratification and severity classification, where rare but clinically important categories are frequently underdetected (Thapa et al., 2025; Baral, Lama, et al., 2025; Lee et al., 2022; Hwang et al., 2022). Evaluating weighting-based ordinal models across these settings will help determine whether gains in minority-class recognition translate into meaningful practical utility under real-world decision constraints.

5. Conclusions

While low grades are increasingly rare, they remain important outcome categories for educational decision-making and early-risk identification. Standard ordinal logistic regression showed limited ability to recognize these minority outcomes, reflecting the challenges posed by imbalanced grade distributions. Incorporating inverse-frequency weighting improved detection of rare grade categories, particularly at the lower end of the distribution, but these gains came at the expense of weaker performance in majority categories and reduced aggregate discrimination. These findings highlight a fundamental trade-off in imbalance-aware ordinal predictive modeling between sensitivity to rare outcomes and overall predictive reliability. Weighted ordinal logistic regression therefore offers a transparent and interpretable approach when minority-outcome detection is prioritized, although richer predictor sets and calibrated or hybrid modeling strategies will be necessary to achieve stronger and more practically useful predictive performance.

Author Contributions: "Conceptualization, D.B[1]. ; methodology, D.B.[1] and D.B.[2] ; software, D.B.[2]; validation, D.B.[1], D.L., and D.B.[2] ; formal analysis, D.B.[1], D.L., and D.B.[2] ; investigation, D.B.[1], D.L., and D.B.[2]; resources, P.K; data curation, D.B.[1], D.L., D.B.[2], and P.K.; writing—original draft preparation, D.B.[1] , and D.B.[2]; writing—review and editing, D.B.[1] , and D.B.[2]; visualization, P.K.; supervision, D.B.[1]; project administration, D.B.[1]. All authors have read and agreed to the published version of the manuscript."

Funding: This research received no external funding.

Institutional Review Board Statement: The study utilized a previously collected, de-identified dataset obtained from Baral et al. (2025). Ethical approval for the original data collection was obtained as reported in the original publication.

Acknowledgments: The authors thank the participants of the original study and the corresponding author for providing access to the dataset. The authors used generative AI tools

(ChatGPT, Grammarly) to assist with language editing during the manuscript preparation process. All ideas, interpretations, and findings are the original work of the authors

Conflicts of Interest: The authors declare no conflicts of interest.

References

Abdulazeez, Y., & Abdulwahab, L. (2018). Application of classification models to predict students' academic performance using classifiers ensemble and synthetic minority over sampling techniques. *Bayero Journal of Pure and Applied Sciences*, *11*(2), 142-148.

Abreu, M. N. S., Siqueira, A. L., Cardoso, C. S., & Caiaffa, W. T. (2008). Ordinal logistic regression models: application in quality of life studies. *Cadernos de Saúde Pública*, *24*, s581-s591.

Achen, A. C., & Courant, P. N. (2009). What are grades made of?. *Journal of Economic Perspectives*, *23*(3), 77-92.

Adejumo, A. O., & Adetunji, A. A. (2013). Application of ordinal logistic regression in the study of students' performance. *Mathematical Theory and Modeling*, *3*(11), 10-19.

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques. *Education Sciences*, *11*(9), 552. https://doi.org/10.3390/educsci11090552

Almalawi, A., Soh, B., Li, A., & Samra, H. (2024). Predictive Models for Educational Purposes: A Systematic Review. *Big Data and Cognitive Computing*, *8*(12), 187.

Arthurs, N., Stenhaug, B., Karayev, S., & Piech, C. (2019). Grades Are Not Normal: Improving Exam Score Models Using the Logit-Normal Distribution. *International Educational Data Mining Society*.

Baral, D. (2025). Leveraging AI, magic, and field trips to prevent math anxiety in elementary school pupils. *American Journal of STEM Education, 9,* 21–34.

Baral, D., Dasgupta, N., McLeman, L., Bista, R., Jones, J., & Biswakarma, M. (2025). Exploring the predictive relationship between students' math anxiety and their mathematical performance. *American Journal of STEM Education, 3,* 1–15.

Baral, D., Lama, D., Dangal, A., Baral, D., Kharel, J., Thapa, A., Dahal, P., & Baral, S. (2025). Ranking Predictors of Child Dietary Diversity in Nepal Using a Decision Tree. *Journal of Public Health Concern Trust (JPHECT), 2*(1), 1–6.

Barroso, C., Ganley, C. M., McGraw, a. L., Geer, E. A., Hart, S. A., & Daucourt, M. C. *(2021)*. A meta-analysis of the relation between math anxiety and math achievement. *Psychological Bulletin, 147(2), 134–168*. https://doi.org/10.1037/bul0000307

Bista, R., Poudel, P., & Baral, D. (2025). Examining the influence of demographic variables on mathematics test anxiety. *American Journal of STEM Education, 3,* 105–119.

Bujang, S. D. A., Selamat, A., Krejcar, O., Mohamed, F., Cheng, L. K., Chiu, P. C., & Fujita, H. (2022). Imbalanced classification methods for student grade prediction: A systematic literature review. *IEEE Access*, *11*, 1970-1989.

Carophine, N. (2016). Ordinal logistic regression model on the factors affecting students perfomance in statistics: A case study of Maasai Mara University.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Ciftci, S. K., & Karadag, E. (2024). Grade inflation effects of capacity expansion in higher education: a longitudinal study in undergraduate teacher education programs from 2003 to 2022. *Humanities and Social Sciences Communications*, *11*(1), 1-9.

Dadzie, J., Anaan-Brew, R. K., & Danso, E. (2025). Gamification in Ghanaian mathematics classrooms: Examining the roles of teacher knowledge, motivation, and self-Efficacy on student engagement. *American Journal of STEM Education*, *15*, 45-86.

Gershenfeld, S., Ward Hood, D., & Zhan, M. (2015). The Role of First-Semester GPA in Predicting Graduation Rates of Underrepresented Students. *Journal of College Student Retention: Research, Theory & Practice*, *17*(4), 469-488. https://doi.org/10.1177/1521025115579251 (Original work published 2016)

Gibson, L. (2026, May 22). *Harvard faculty approve a cap on a grades*. Harvard Magazine. https://www.harvardmagazine.com/university-news/harvard-faculty-approve-a-cap-on-a-grades

Hanford, E. (2016). Stuck at square one: College students increasingly caught in remedial education trap. *APM Reports*.

Husong, J. P. (2024). *"Bouncing Back": The Development, Implementation, and Effectiveness of an Academic Probation Support Program*. University of Dayton.

Hwang, J. Y., Lee, C., Nguyen, T., Buck, T., Mun, K., Vanteru, I. T., Baral, D., Tirschwell, D., & Kim, A. (2022). Characteristics and outcomes of hospitalized COVID-19 patients and DNR. *Journal of Investigative Medicine, 70*(1), 267. BMJ Publishing Group.

Jenkins, D., Jaggars, S. S., & Roksa, J. (2009). Promoting Gatekeeper Course Success among Community College Students Needing Remediation: Findings and Recommendations from a Virginia Study (Summary Report). *Community College Research Center, Columbia University*.

Jeong, H., Weiss, G. M., Leung, a., & Leeds, d. D. *(2024)*. The construction and analysis of course grades across public universities. *In Proceedings Of The 17Th International Conference On Educational Data Mining (Pp*. 1-4). *International Educational Data Mining Society*. https://doi.org/10.5281/zenodo.12729906

Jephcote, C., Medland, E., & Lygo-Baker, S. (2021). Grade inflation versus grade improvement: Are our students getting more intelligent?. *Assessment & Evaluation in Higher Education*, *46*(4), 547-571.

Jewell, R. T., McPherson, M. A., & Tieslau, M. A. (2013). Whose fault is it? Assigning blame for grade inflation in higher education. *Applied economics*, *45*(9), 1185-1200.

Kamber, R. (2008). Understanding grade inflation. *Grade inflation: Academic standards in higher education*, 45-71.

Karadağ, E. *(2021)*. Effect of COVID-19 pandemic on grade inflation in higher education in Turkey. *PLOS ONE, 16(8), e0256680*. https://doi.org/10.1371/journal.pone.0256680

Kearney, C. A., Fensken, M., & Dupont, R. (2024, January). The K-16 education movement: common themes across K-12 and higher education systems to inform development and evaluation. In *Frontiers in Education* (Vol. 9, p. 1272297). Frontiers Media SA.

Lee, C., Nguyen, T., Hwang, J. Y., Buck, T., Mun, K., Vanteru, I. T., Baral, D., Tirschwell, D., & Kim, A. (2022). Therapeutic anticoagulation in hospitalized patients with COVID-19. *Journal of Investigative Medicine, 70*(1), 268–269. BMJ Publishing Group.

1. Levine, A., & Cureton, J. S. (1998). *When hope and fear collide: A portrait of today's college student*. San Francisco, CA: Jossey-Bass Publishers

Limanto, S., Buliali, J. L., & Saikhu, A. (2024). GLoW SMOTE-D: Oversampling technique to improve prediction model performance of students failure in courses. *IEEE Access*, *12*, 8889-8901.

Liu, X., & Whitford, M. *(2011)*. Opportunities to learn in school and mathematics proficiency: A continuation ratio model. *Journal Of Educational Research, 104(5), 328–339.*

Martinez, J. A., Sher, K. J., Krull, J. L., & Wood, P. K. (2009). Blue-Collar Scholars?: Mediators and Moderators of University Attrition in First-Generation College Students. *Journal of college student development*, *50*(1), 87–103. https://doi.org/10.1353/csd.0.0053

Muya, K. B. (2018). *Application of Ordinal Logistic Regression in Analyzing Students' Performance at Kenya Certificate of Secondary Education Level in Kiambu County* (Doctoral dissertation, University of Nairobi).

Namkung, J. M., Peng, p., & Lin, x. *(2019)*. The relation between mathematics anxiety and mathematics performance among school-aged learners: A meta-analysis. *Review Of Educational Research, 89(3), 459–496.*

Pattison, E., Grodsky, E., & Muller, C. (2013). Is the Sky Falling? Grade Inflation and the Signaling Power of Grades. *Educational Researcher*, *42*(5), 259-265. https://doi.org/10.3102/0013189X13481382

Pearce II, J. A. (2017). How employers can stanch the hemorrhaging of collegiate GPA credibility. *Business Horizons*, *60*(1), 35-43.

Polyzou, A., & Karypis, G. *(2019)*. Feature extraction and grade prediction in large-scale courses using machine learning. *Proceedings Of The 12Th International Conference On Educational Data Mining, 261–270.*

Ramirez, G., Shaw, S. T., & Maloney, e. A. *(2019)*. Math anxiety: Past research, promising interventions, and a new interpretation framework. *Educational Psychologist, 53(3), 145–164.*

Ran, F. X., & Lin, Y. (2022). Rethinking remedial programs to promote college student success. *Brookings Institution*, *15*.

Rodríguez-Planas, N. (2022). COVID-19, college academic performance, and the flexible grading policy: A longitudinal analysis. *Journal of Public Economics*, *207*, 104606.

Rojstaczer, S., & Healy, C. (2012). Where A is ordinary: The evolution of American college and university grading, 1940-2009. *Teachers College Record*, *114*(7), 1-23.

Saa, A. A. (2016). Educational data mining & students' performance prediction. *International journal of advanced computer science and applications*, *7*(5).

Sanabria, T., Penner, A., & Domina, T. (2020). Failing at remediation? College remedial coursetaking, failure and long-term student outcomes. *Research in Higher Education*, *61*(4), 459-484.

Saputro, D. R. S., & Widyaningsih, P. (2017, August). Limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method for the parameter estimation on geographically weighted ordinal logistic regression model (GWOLR). In *AIP conference proceedings* (Vol. 1868, No. 1, p. 040009). AIP Publishing LLC.

Sesay, R. B., Kpangay, M., & Seppeh, S. (2021). An ordinal logistic regression model to identify factors influencing students academic performance at Njala University. *Int. J. Res. Sci. Innov*, *8*(01), 91-100.

Thapa, A., Baral, D., Dahal, P., Kharel, J., Dangal, A., Baral, S., Thapa, R., Paudel, P., & Khadka, R.. (2025). Dietary diversity and its associated factors among children aged 6–59 months in Madhyapur Thimi Municipality, Nepal. *Nepal Journal of Public Health, 2*(1), 50–56.

Tucker, J., & Courts, B. (2010). Grade inflation in the college classroom. *Foresight*, *12*(1), 45-53.

Walters, G. D., Kremser, j., & Runck, b. *(2018)*. Predicting community college outcomes: A comparison of ordinal and linear models. *Community College Journal Of Research And Practice, 42(10), 684–699*. https://doi.org/10.1080/10668926.2017.1352685

Wongvorachan, T., He, S., & Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information*, *14*(1), 54. https://doi.org/10.3390/info14010054

Zhou, F., Wu, D., Yang, X., & Jiao, J. (2008, December). Ordinal logistic regression for affective product design. In *2008 IEEE International Conference on Industrial Engineering and Engineering Management* (pp. 1986-1990). IEEE.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the publisher and/or the editor(s). The publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.