Early Grade Prediction and Validation to Support Students in a Foundational STEM Course

Thomas Joyce, Alison Cheng, Ph.D., G. Alex Ambrose, Ph.D., Shawn Miller, Ph.D., & Bo Pei, Ph.D.

UNIVERSITY OF NOTRE DAME

Background: Learning Analytics

- **Learning analytics** is the measurement, collection, analysis, and reporting of data about learners and their contexts, for the purposes of understanding and optimizing learning and the environment in which it occurs[1]
- Learning data comes in many forms: Grade histories from learning management systems, clickstream data from e-learning platforms, student survey data
- Key goals of learning analytics
  - Provide personalized and timely learning feedback to students
  - Help instructors monitor student performance and develop effective teaching strategies
  - Use early grade prediction to identify at-risk students

[1](“What Is Learning Analytics,” *SoLAR*)

Background: Early Grade Prediction

- **Early grade prediction** – Prediction of students’ final grades early in the semester to help instructors identify students who are at risk of failing or dropping out of a course
- Use limited data to make predictions
- Outcome variable is usually binary (pass/fail)
- Several studies have employed machine learning methods to accurately predict students’ academic grades at early course stages (Al-Shabandar et al., 2019; Marbouti et al., 2016; Riestra-González et al., 2021)
- Common approach is to compare different ML methods (e.g., Logistic Regression, Random Forest, Gradient Boosting, Naive Bayes) for identifying at-risk students
- Less emphasis on validation of model performance

Should we use demographic data as predictor variables?

- Common demographic variables in student risk prediction: race/ethnicity, sex/gender, disabilities, free or reduced meal price availability, English as a second language status[2]
- Arguments for using demographic data as predictors
    - May lead to more accurate predictions of at-risk students[2]
- Arguments against using demographic data as predictors
    - Might suppress actionable variables[2]
    - Could reinforce biases in the training labels and be harmful to historically underrepresented groups[2]
    - “Fairness through unawareness”: Algorithm is fair if it does not see protected attributes in decision-making process[3]
- A middle ground?
    - Use demographic variables to apply fairness constraints in model without explicitly including them as predictors[4]

[2](Baker et al., 2023), [3](Kusner et al., 2017), [4](Zafar et al., 2019)

Our Study

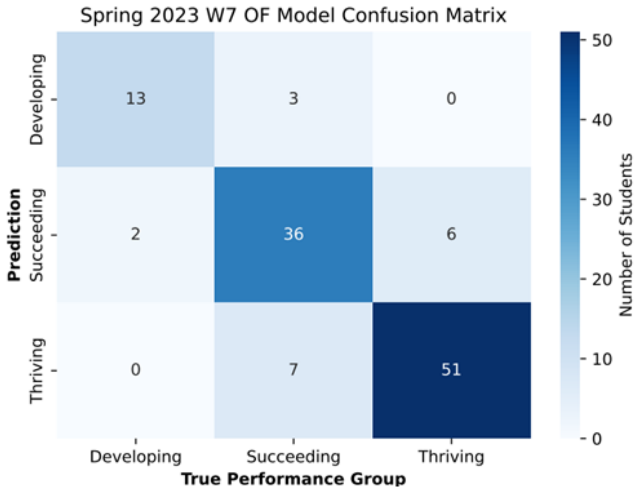
- “Early Grade Prediction and Validation to Support Students in a Foundational STEM Course”
- **STEM Course**: Organic Chemistry at a mid-sized private university
- **Early Grade Prediction**: Predict students’ final performance groups before the midterm break
    - No demographic data – We did not have access to this
- **Validation**: We used Spring 2023 course data to train our models, and now we want to test them on the Spring 2024 data
    - Validation is essential if course instructors want to use our models to identify at-risk students and deliver timely learning interventions

Previous Findings

- Created ordinal forest models to predict students' final performance groups each week before the midterm break in the Spring 2023 semester
- Highly accurate grade predictions by Week 7
- **Will these results hold in Spring 2024?**

Week 7 Ordinal Forest Model Results Spring 2023 (Performance Group ~ Quiz 1 Attempt 1 + Quiz 2 Attempt 1 + Quiz 3 Attempt 1 + Exam 1 + Quiz 4 Attempt 1 + Quiz 5 Attempt 1 + Exam 2)

Metric	Value
Accuracy	0.8475
Quadratic Weighted Kappa	0.84
Non-Thriving Sensitivity	0.8852



Course Overview

- This study focuses on the first course of a two-semester sequence in organic chemistry for science students
- Students usually take this course in the spring semester of their first year
- Course meets three times per week for 50-minute lectures
- Spring 2023 enrollment: 391 students
- Spring 2024 enrollment: 423 students
- Data source: Canvas gradebook

Grading Scheme

Course Assessments

Assessment	Percentage of Course Grade in Spring 2023	Percentage of Course Grade in Spring 2024
Canvas Quizzes* (10)	5.5%	6.8%
Tutorials (13)	9.1%	9.0%
Midterm Exams (4)	61.0%	60.2%
Final Exam (1)	24.4%	24.0%

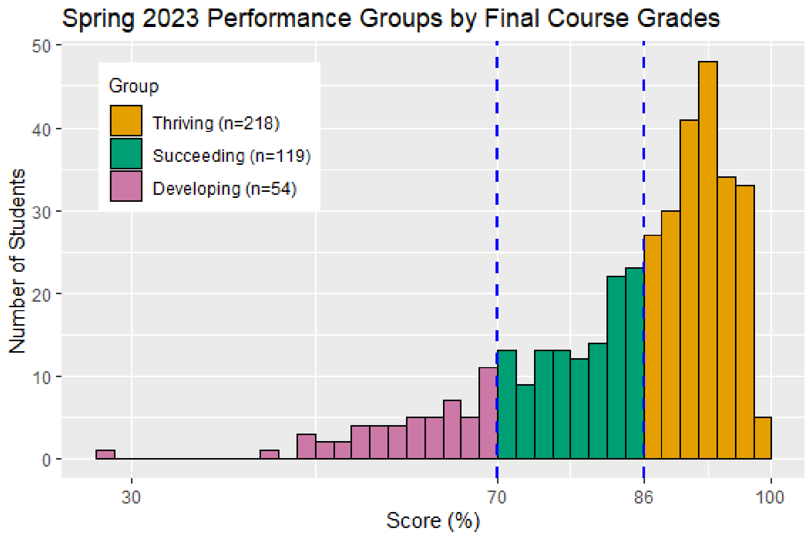
*Students had two identical attempts on each Canvas Quiz and the highest score was accepted

Letter Grade by Final Percentage

Final Course Grade Percentage	Letter Grade
90-100	A
86-89	A-
83-85	B+
80-82	B
75-79	B-
70-74	C+
65-69	C
60-64	C-
50-59	D
<50	F

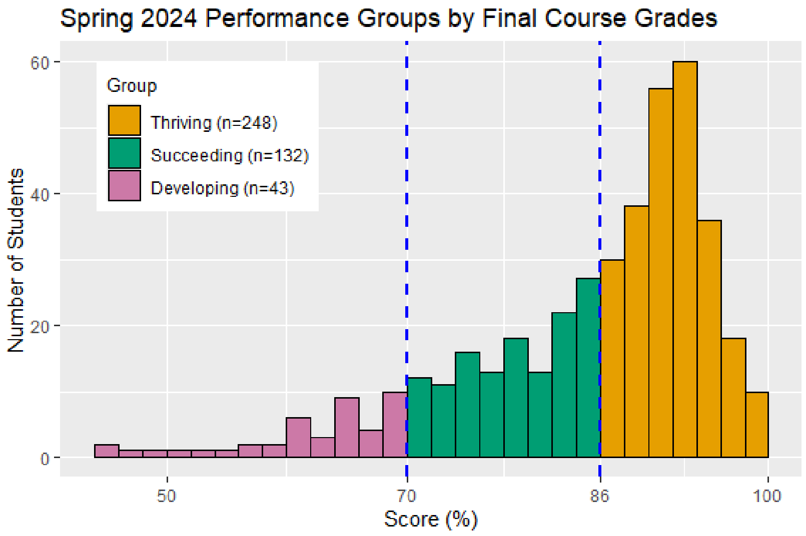
Defining the Performance Groups

Spring 2023



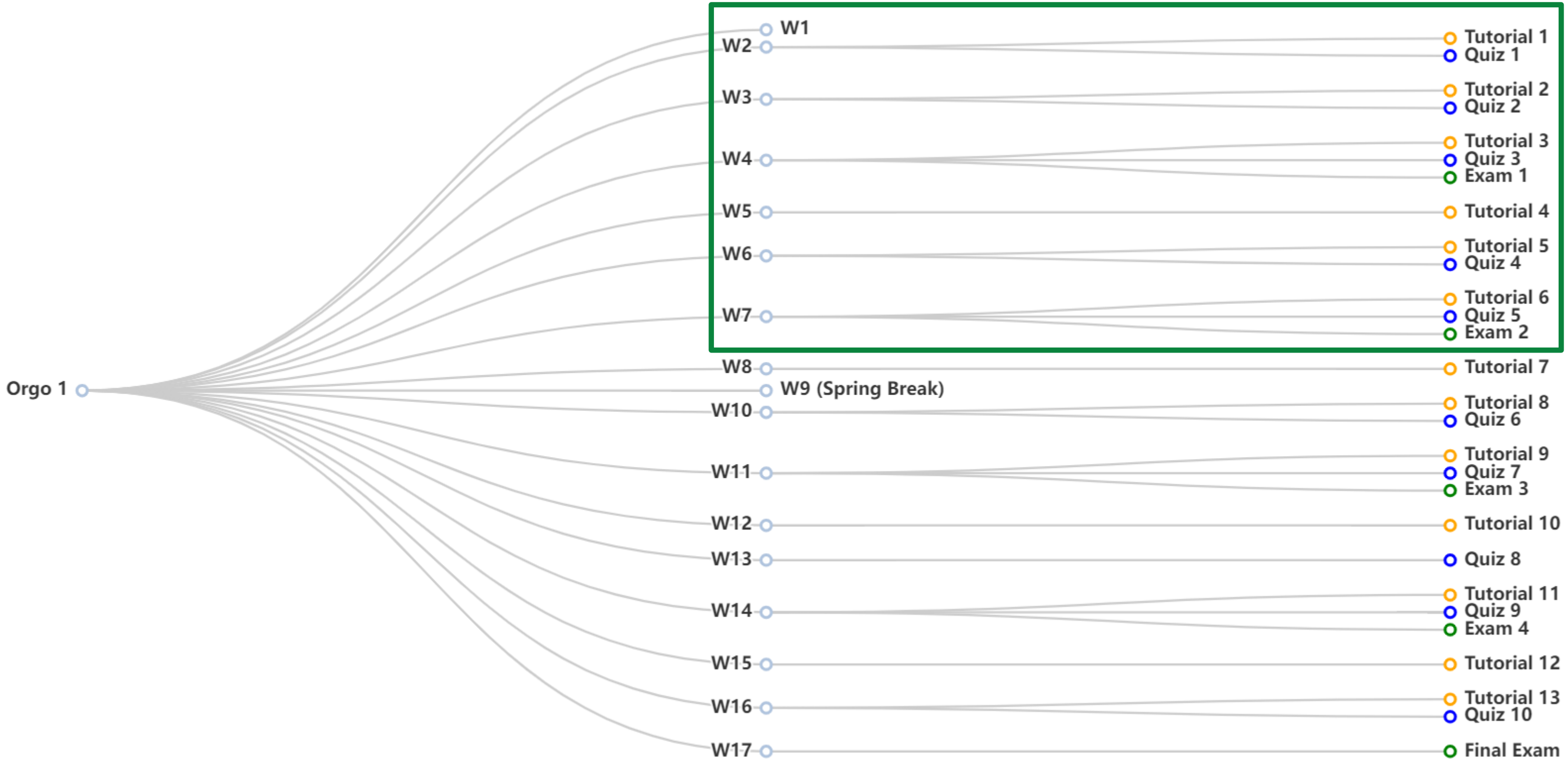
Group	Letter Grades	Number of Students	Percent of Students
Thriving	A or A-	218	55.8%
Succeeding	B+, B, B-, C+	119	30.4%
Developing	C or below	54	13.8%

Spring 2024



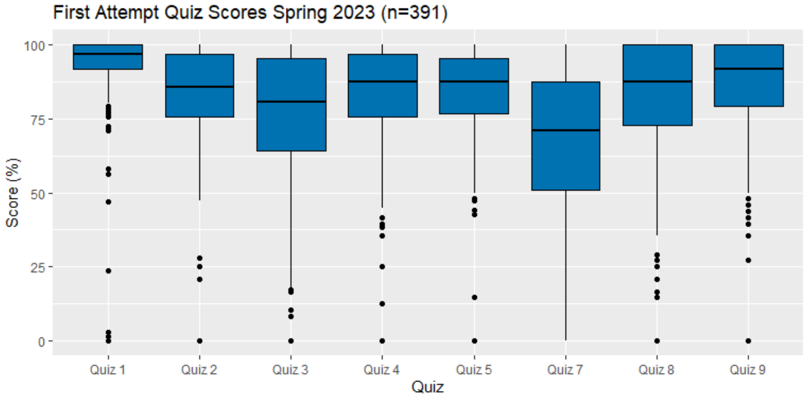
Group	Letter Grades	Number of Students	Percent of Students
Thriving	A or A-	248	58.6%
Succeeding	B+, B, B-, C+	132	31.2%
Developing	C or below	43	10.2%

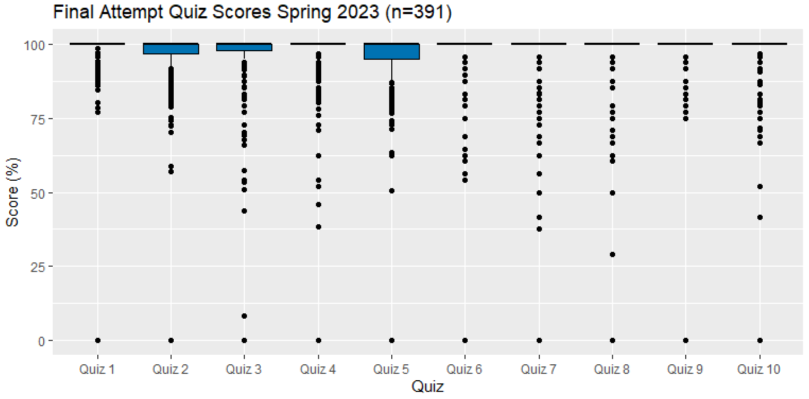
Course Assessment Timeline



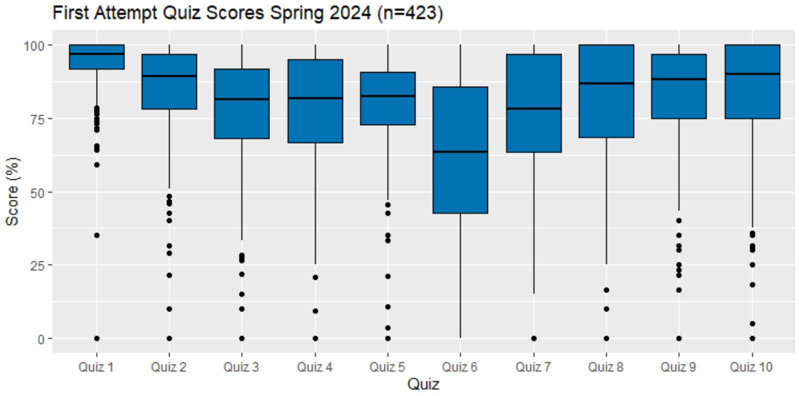
Higher Variation on First Attempt Quiz Scores

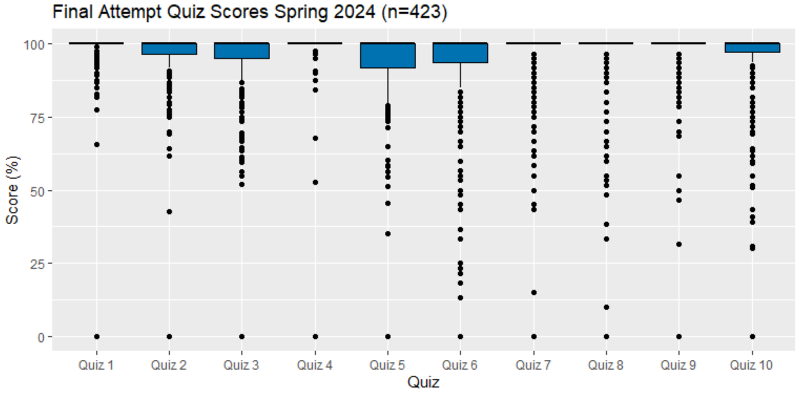
Spring 2023





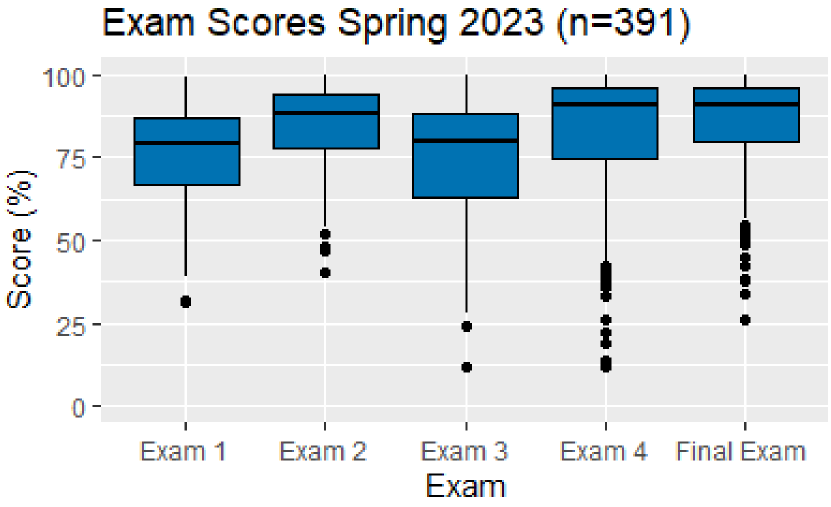
Spring 2024



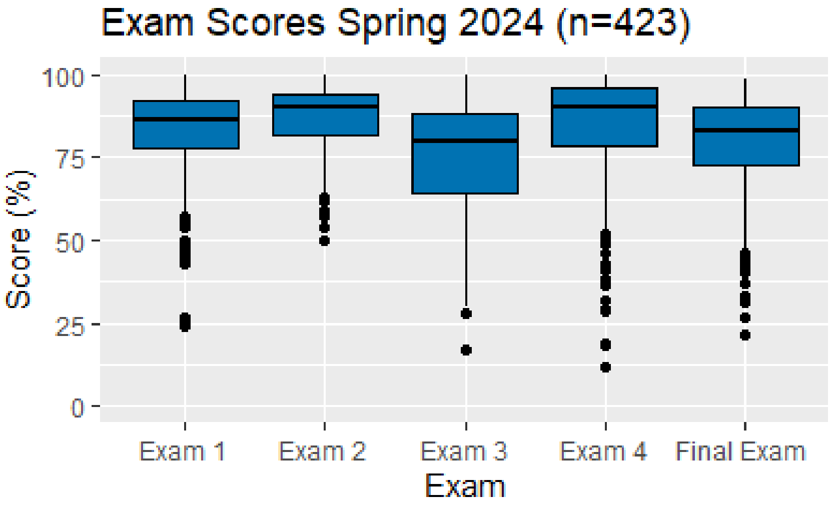


Similar Exam Score Distributions between Semesters

Spring 2023



Spring 2024



Research Questions

- 1) Is it possible to create models that accurately predict students' final performance groups at early course stages without using demographic variables?
- 2) Can we replicate the Spring 2023 predictive model results in Spring 2024?

Methods

- **Ordinal forest (OF)** models to predict final performance groups (Thriving, Succeeding, Developing) in Weeks 3-7
- OF is a random forest-based method for ordinal response variables[5]
- By taking the ordinal nature of the response variable into account, OF yields fewer predictions that are far from the true class value[5]
- Predictors: All first attempt quiz scores and exam scores leading up to the specified week
- 70% of the Spring 2023 data for the training set and 30% of the Spring 2023 data for the test set
- Balanced the training data using SMOTE
- Used the ordinalForest package in R to train the models
- Test Spring 2023 models on Spring 2024 data

[5](Hornung, 2020)

Model Evaluation Metrics

- **Accuracy**
  - How well did the classification models perform overall?
- **Quadratic Weighted Kappa**[6]
  - Measure of agreement between predictions and true labels
  - Quadratic weights give a larger penalty for predictions that are farther from the true class value
  - Kappa scale: 0.01 – 0.20 (Slight agreement), 0.21 – 0.40 (Fair agreement), 0.41 – 0.60 (Moderate agreement), 0.61 – 0.80 (Substantial agreement), 0.81 – 1.0 (Almost perfect agreement)
- **Non-Thriving Sensitivity**
  - $\text{Non}-\text{thriving sensitivity} = \dfrac{\text{Number of students correctly classifed as non}-\text{thriving}}{\text{Number of students who are truly non}-\text{thriving}}$
  - Here, non-thriving means Succeeding or Developing
  - Non-thriving students are more likely to be at risk of failing or dropping out of the course

[6](Cohen, 1968)

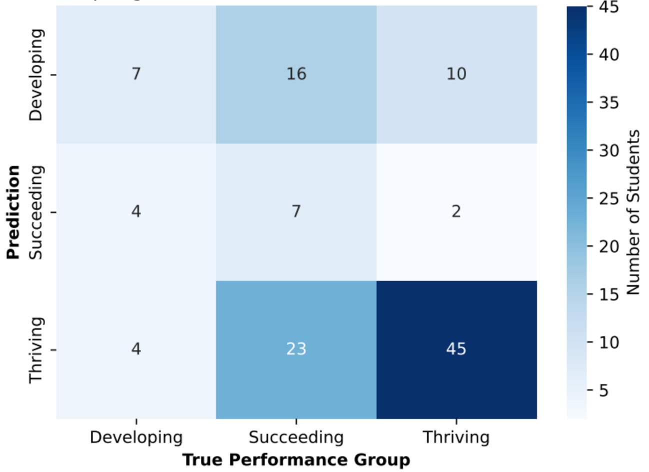
Classification Models Results

Classification Models Results Week 3

Spring 2023

Week 3 Ordinal Forest Model Results Spring 2023 (Performance Group ~ Quiz 1 Attempt 1 + Quiz 2 Attempt 1)

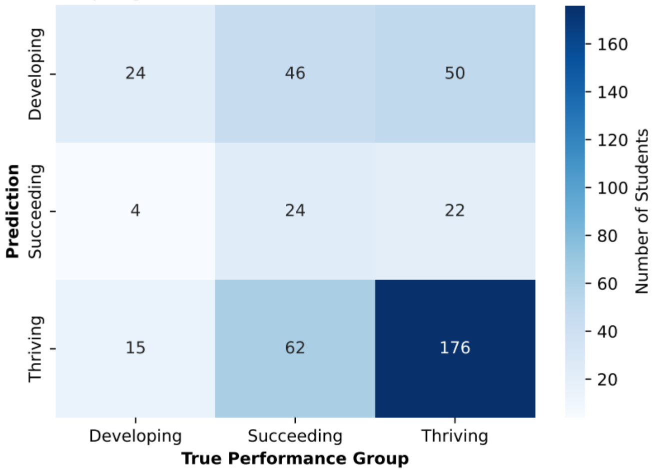
Metric	Value
Accuracy	0.5
Quadratic Weighted Kappa	0.32
Non-Thriving Sensitivity	0.5574



Spring 2024

Week 3 Ordinal Forest Model Results Spring 2024 (Performance Group ~ Quiz 1 Attempt 1 + Quiz 2 Attempt 1)

Metric	Value
Accuracy	0.5296
Quadratic Weighted Kappa	0.26
Non-Thriving Sensitivity	0.56

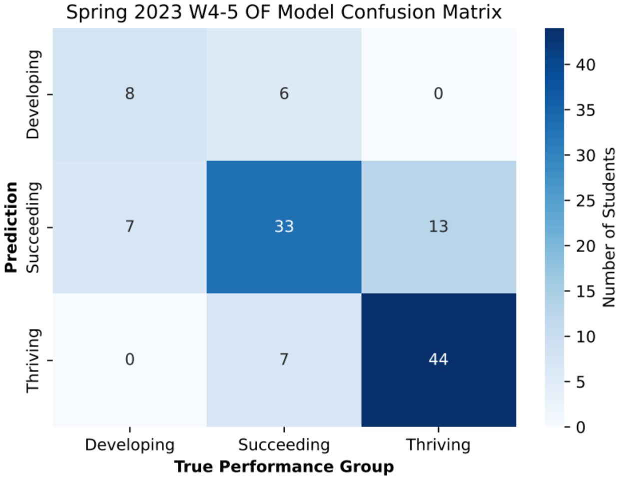


Classification Models Results Weeks 4-5

Spring 2023

Weeks 4-5 Ordinal Forest Model Results Spring 2023 (Performance Group ~ Quiz 1 Attempt 1 + Quiz 2 Attempt 1 + Quiz 3 Attempt 1 + Exam 1)

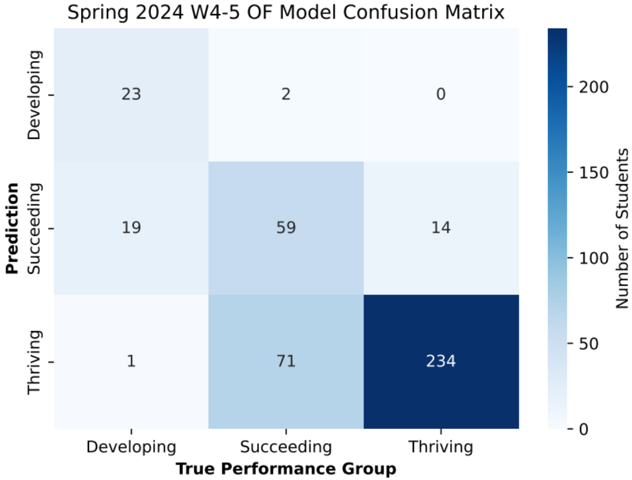
Metric	Value
Accuracy	0.7203
Quadratic Weighted Kappa	0.70
Non-Thriving Sensitivity	0.8852



Spring 2024

Weeks 4-5 Ordinal Forest Model Results Spring 2024 (Performance Group ~ Quiz 1 Attempt 1 + Quiz 2 Attempt 1 + Quiz 3 Attempt 1 + Exam 1)

Metric	Value
Accuracy	0.747
Quadratic Weighted Kappa	0.69
Non-Thriving Sensitivity	0.5886

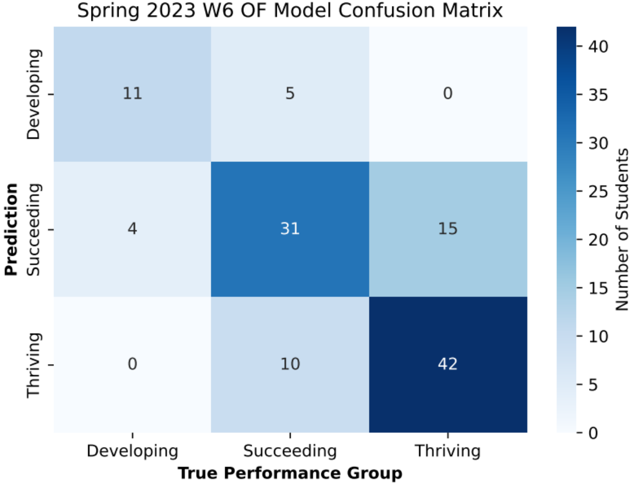


Classification Models Results Week 6

Spring 2023

Week 6 Ordinal Forest Model Results Spring 2023 (Performance Group ~ Quiz 1 Attempt 1 + Quiz 2 Attempt 1 + Quiz 3 Attempt 1 + Exam 1 + Quiz 4 Attempt 1)

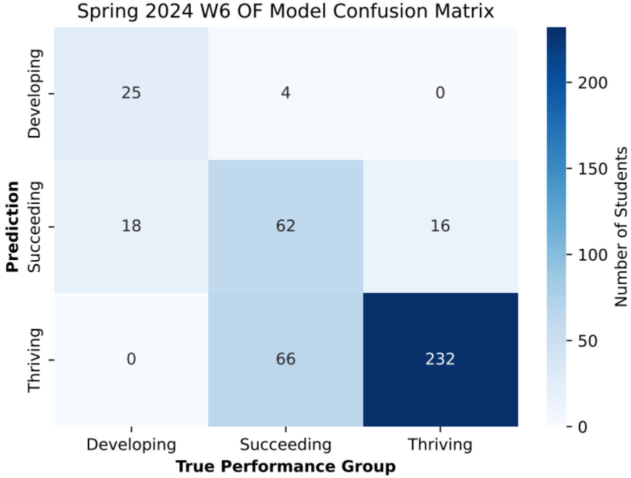
Metric	Value
Accuracy	0.7119
Quadratic Weighted Kappa	0.70
Non-Thriving Sensitivity	0.8361



Spring 2024

Week 6 Ordinal Forest Model Results Spring 2024 (Performance Group ~ Quiz 1 Attempt 1 + Quiz 2 Attempt 1 + Quiz 3 Attempt 1 + Exam 1 + Quiz 4 Attempt 1)

Metric	Value
Accuracy	0.7541
Quadratic Weighted Kappa	0.71
Non-Thriving Sensitivity	0.6229

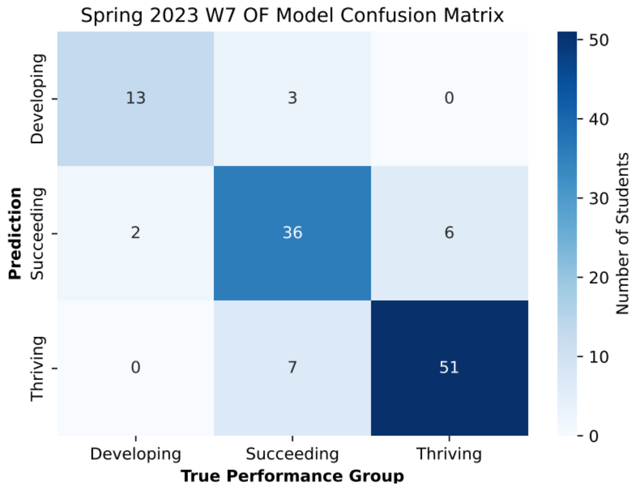


Classification Models Results Week 7

Spring 2023

Week 7 Ordinal Forest Model Results Spring 2023 (Performance Group ~ Quiz 1 Attempt 1 + Quiz 2 Attempt 1 + Quiz 3 Attempt 1 + Exam 1 + Quiz 4 Attempt 1 + Quiz 5 Attempt 1 + Exam 2)

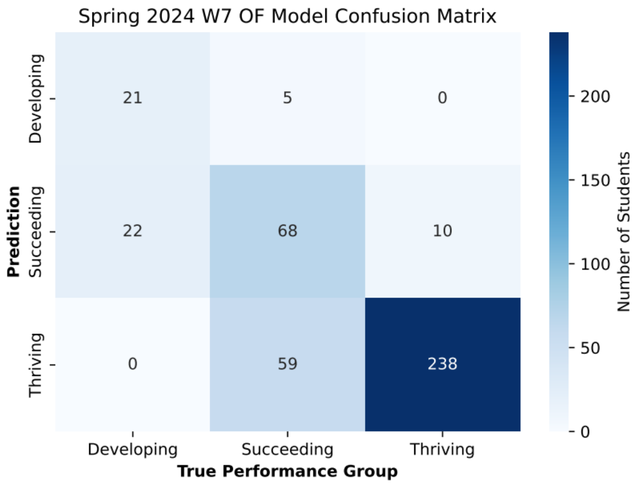
Metric	Value
Accuracy	0.8475
Quadratic Weighted Kappa	0.84
Non-Thriving Sensitivity	0.8852



Spring 2024

Week 7 Ordinal Forest Model Results Spring 2024 (Performance Group ~ Quiz 1 Attempt 1 + Quiz 2 Attempt 1 + Quiz 3 Attempt 1 + Exam 1 + Quiz 4 Attempt 1 + Quiz 5 Attempt 1 + Exam 2)

Metric	Value
Accuracy	0.773
Quadratic Weighted Kappa	0.73
Non-Thriving Sensitivity	0.6629



Discussion

- Spring 2023: Accuracy increased each week. Good predictions by Week 4 and very accurate predictions by Week 7
- Spring 2024: Achieved similar overall accuracy and quadratic weighted Kappa in Weeks 4-7 but lower non-thriving sensitivity
- Satisfactory early grade predictions without using demographic data
- Results were nearly replicable in Spring 2024 but could be better

- Our study follows an analogous approach to training-testing-cross validation in traditional machine learning: 70% of Spring 2023 data used for training, 30% of Spring 2023 data used for hold-out cross-validation, and Spring 2024 data used for testing
- Potential improvement: Use stratified train-test split on the Spring 2023 data to preserve original performance group distributions

Percentage of Students in each Performance Group

Spring 2023 Entire Course (n=391)

Group	Letter Grades	Number of Students	Percent of Students
Thriving	A or A-	218	55.8%
Succeeding	B+, B, B-, C+	119	30.4%
Developing	C or below	54	13.8%

30%

70%

Spring 2023 Test Set (n=118)

Group	Letter Grades	Number of Students	Percent of Students
Thriving	A or A-	57	48.3%
Succeeding	B+, B, B-, C+	46	39.0%
Developing	C or below	15	12.7%

Spring 2023 Training Set (n=273)

Group	Letter Grades	Number of Students	Percent of Students
Thriving	A or A-	161	59.0%
Succeeding	B+, B, B-, C+	73	26.7%
Developing	C or below	39	14.3%

Spring 2024 Entire Course (n=423)

Group	Letter Grades	Number of Students	Percent of Students
Thriving	A or A-	248	58.6%
Succeeding	B+, B, B-, C+	132	31.2%
Developing	C or below	43	10.2%

Future Directions

- Explore other machine learning methods and compare them to ordinal forest
- Continue to test and refine the models in subsequent semesters
- Item Analysis
    - Compare performance groups on specific exam topics and questions to provide instructors with more detailed insights
- Assessment Wrappers
    - Short post-exam reflection surveys that ask students questions about their exam preparation
    - Use wrapper data to evaluate individual fairness of our models: Did students with similar wrapper responses receive similar performance group predictions from the models?

How can we use assessment wrappers?

Assessment Wrapper Prompts

Preparation: How prepared did you feel coming into this quiz/exam? Please select your answer using the ratings in the table below.
(Extremely prepared, very prepared, somewhat prepared, a little prepared, completely unprepared)

Confidence: How do you feel about your performance on the quiz/exam? Please select your answer using the ratings in the table below.
(Extremely confident, very confident, somewhat confident, a little confident, not confident

Content: What topics in the quiz/exam did you feel unprepared to answer, if any? Using keywords (i.e. [examples]) can help you think about your answer.

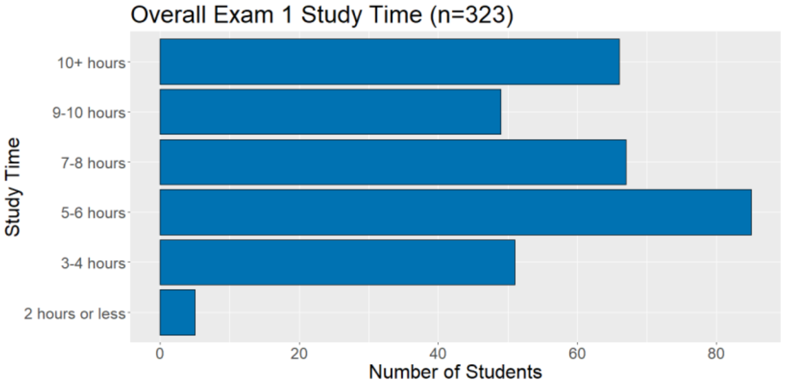
Study Total Time: How much time did you spend preparing for this quiz/exam?
(None, less than one hour, 1-2 hours, 3-4 hours, 5-6 hours, 7-8 hours, 9-10 hours, 10+ hours)

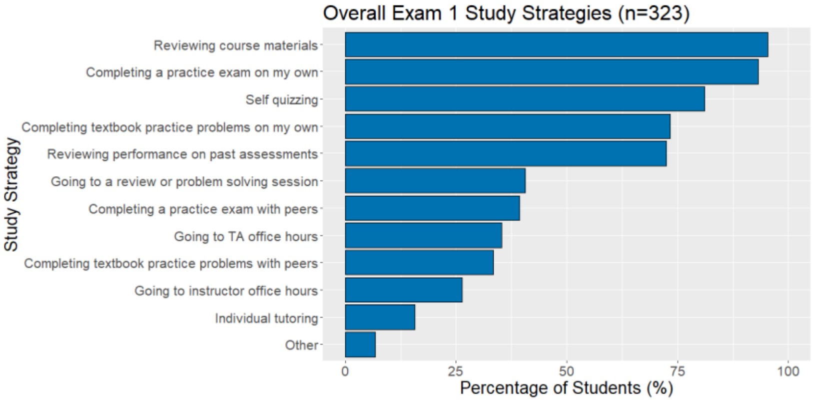
Study Strategies: What tools and strategies did you use to prepare for this quiz/exam?
Check all that apply.

- reviewing course materials (i.e. lectures, notes, etc)
- going to a review or problem solving session,
- self quizzing (i.e. flash cards)
- completing textbook practice problems on my own
- completing textbook practice problems with peers
- completing a practice exam on my own
- completing a practice exam with peers
- reviewing performance on past assessments (quizzes/Tutorial problem sets)
- going to instructor office hours
- going to TA office hours
- individual tutoring
- other (please select this and then answer the next question with specifics)

If you answered "other" in the previous question, then please specify below what other tool or strategy you used to prepare for this exam. If you did not select "other", please write "N/A".

Stress: On a scale of 1 - 5, indicate your level of stress while preparing for this quiz/exam? Please select your answer using the ratings in the table below.
(Extremely stressed, very stressed, somewhat stressed, a little stressed, not at all stressed)





References

- Al-Shabandar, R., Hussain, A. J., Liatsis, P., & Keight, R. (2019). Detecting At-Risk Students With Early Interventions Using Machine Learning Techniques. *IEEE Access*, *7*, 149464–149478. https://doi.org/10.1109/ACCESS.2019.2943351
- Baker, R. S., Esbenshade, L., Vitale, J., & Karumbaiah, S. (2023). Using Demographic Data as Predictor Variables: A Questionable Choice. *Journal of Educational Data Mining*, *15*(2), 22–52.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin,* 70(4), 213.
- Hornung, R. (2020). Ordinal Forests. *Journal of Classification*, *37*(1), 4–17. https://doi.org/10.1007/s00357-018- 9302-x
- Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html
- Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, *103*, 1–15. https://doi.org/10.1016/j.compedu.2016.09.005
- Riestra-González, M., Paule-Ruíz, M. del P., & Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Computers & Education*, *163*, 104108. https://doi.org/10.1016/j.compedu.2020.104108
- What is Learning Analytics. (n.d.). *Society for Learning Analytics Research (SoLAR)*. Retrieved June 6, 2024, from https://www.solaresearch.org/about/what-is-learning-analytics/
- Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019). Fairness Constraints: A Flexible Approach for Fair Classification. *Journal of Machine Learning Research*, *20*(75), 1–42.

UNIVERSITY OF
NOTRE DAME

Ordinal Forest Models Variable Importance Values

Week 3 Ordinal Forest Variable Importance

Variable	Importance
Quiz 1 Attempt 1	0.20747
Quiz 2 Attempt 1	0.19488

Weeks 4-5 Ordinal Forest Variable Importance

Variable	Importance
Exam 1	0.49293
Quiz 2 Attempt 1	0.04340
Quiz 3 Attempt 1	0.03468
Quiz 1 Attempt 1	0.01988

Ordinal Forest Models Variable Importance Values

Week 6 Ordinal Forest Variable Importance

Variable	Importance
Exam 1	0.44690
Quiz 2 Attempt 1	0.04657
Quiz 3 Attempt 1	0.02903
Quiz 1 Attempt 1	0.02362
Quiz 4 Attempt 1	0.01788

Week 7 Ordinal Forest Variable Importance

Variable	Importance
Exam 1	0.26771
Exam 2	0.23020
Quiz 2 Attempt 1	0.02297
Quiz 1 Attempt 1	0.01491
Quiz 4 Attempt 1	0.01445
Quiz 3 Attempt 1	0.01390
Quiz 5 Attempt 1	0.01053

Ordinal Forest Algorithm

- "The OF algorithm consists of the following two main steps:

1. **Optimization of the score set:** As described in the "Introduction," ordinal forests are regression forests in which the class values are replaced by score values that are optimized with the aim of maximizing the (OOB) prediction performance. The first step in the optimization of the score set $\{s_1,\ldots,s_J\}$ is performed as follows: First, repeatedly and randomly generate a candidate score set $\{s_{b,1},\ldots,s_{b,J}\}$; second, construct an OF as a regression forest using $\{s_{b,1},\ldots,s_{b,J}\}$ for the class values of the target variable; and lastly, measure the OOB prediction performance according to a specific measure, called the performance function. In the second step, the final score set is calculated as a summary of the score sets that featured the highest OOB prediction performance in the first step.

2. **Construction of the OF as a regression forest:** Using $\{s_1,\ldots,s_J\}$ for the class values of target variable, construct an ordinal forest as a regression forest"

Hornung, R. (2020). Ordinal Forests. *Journal of Classification*, *37*(1), 4–17. https://doi.org/10.1007/s00357-018- 9302-x

Ordinal Forest Model Parameters

- The following default parameters were used to train the OF models using the ordinalForest package in R:
  - nsets = 1000 (Number of score sets tried prior to the approximation of the optimal score set.)
  - ntreeperdiv = 100 (Number of trees in the smaller regression forests constructed for each of the nsets different score sets tried.)
  - ntreefinal = 5000 (Number of trees in the larger regression forest constructed using the optimized score set.)
  - importance = “rps” (The type of variable importance measure to use; the default “rps” uses the ranked probability score as an error measure.)
  - perffunction = “equal” (Performance function; use perffunction = “equal” if it is of interest to classify observations from each class with the same accuracy independent of the class sizes.)

Hornung, R. (2022). _ordinalForest: Ordinal Forests: Prediction and Variable Ranking with Ordinal Target Variables_. R package version 2.4-3, https://CRAN.R-project.org/package=ordinalForest

Hornung, R. (2020). Ordinal Forests. *Journal of Classification*, *37*(1), 4–17. https://doi.org/10.1007/s00357-018- 9302-x

Tutorial Scores

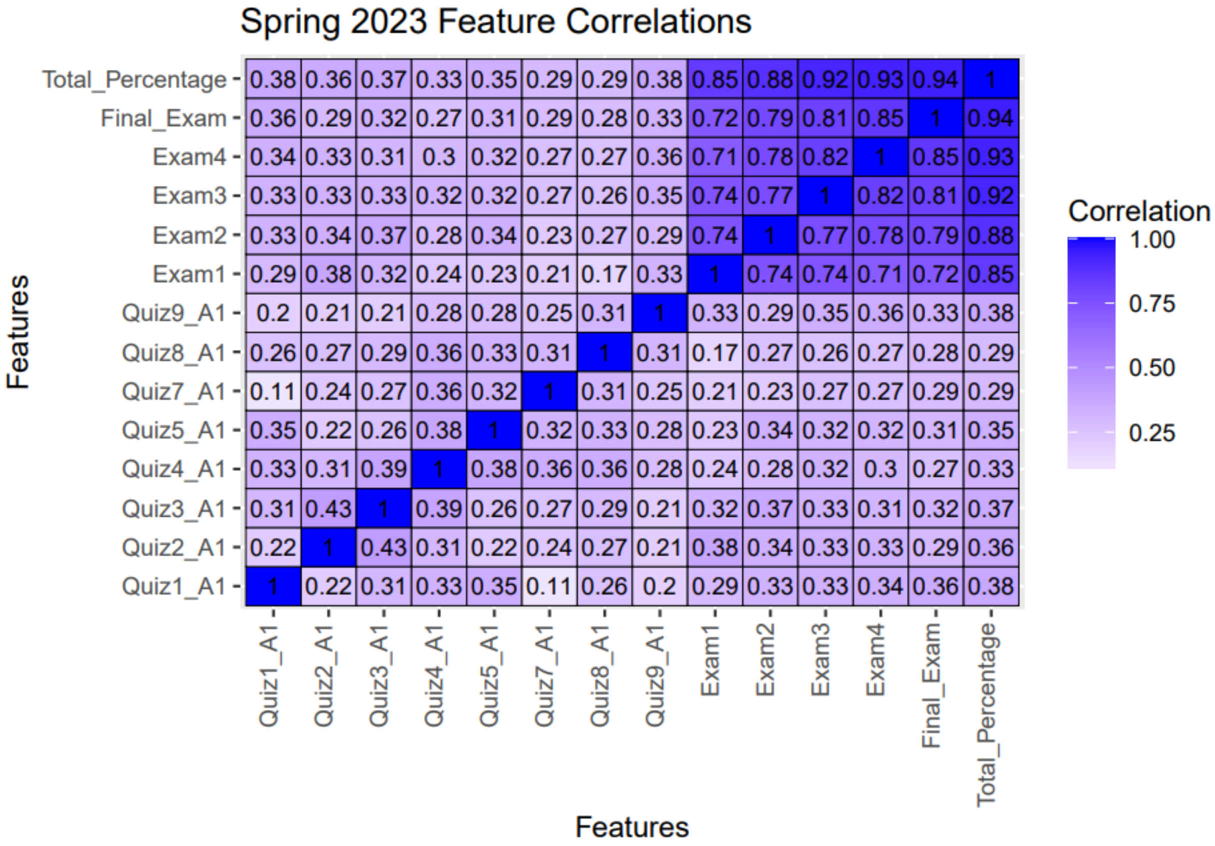
Spring 2023

Tutorial	Number of students who scored 100%	Pct. of students who scored 100%
1	389	99.5%
2	388	99.2%
3	389	99.5%
4	387	99.0%
5	385	98.5%
6	389	99.5%
7	384	98.2%
8	379	97.0%
9	382	97.7%
10	379	97.0%
11	383	98.0%
12	384	98.2%
13	376	96.2%

Spring 2024

Tutorial	Number of students who scored 100%	Pct. of students who scored 100%
1	421	99.5%
2	421	99.5%
3	419	99.1%
4	417	98.6%
5	419	99.1%
6	421	99.5%
7	418	98.8%
8	415	98.1%
9	415	98.1%
10	416	98.3%
11	417	98.6%
12	414	97.9%
13	403	95.3%

Spring 2023 Heatmap for Feature Correlations



Spring 2024 Heatmap for Feature Correlations