JOURNAL OF LEARNING ANALYTICS
SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH
Volume 13(1), 89–109. https://doi.org/10.18608/jla.2026.9099

Modelability as a Strategy for Improving the Generalizability and Scalability of Predictive Models

Alice Xu[1], Icy (Yunyi) Zhang[2], Adam B. Blake[3] and James W. Stigler[4]

Abstract
Learning analytics has the potential to enhance education through data-informed decision-making, but persistent challenges around generalizability and scalability continue to limit its real-world impact. In this paper, we introduce the concept of a modelable world: a learning ecosystem purposefully designed to support the development of predictive models that generalize across diverse contexts. We outline three core design principles of modelability: (1) valid and interpretable measurements, (2) scalable and stable implementation, and (3) a collaborative research–practice–technology ecosystem. We then illustrate how these principles can be operationalized in the real world through a case study of CourseKata, a platform offering a fully instrumented online textbook adopted across a wide range of institutions and disciplines. Using CourseKata data, we developed early prediction models of students' final course grades using behavioral measures and tested the model generalizability across institutions (something rarely done in the modeling literature). Results show that a system designed with modelability in mind can produce predictive models that generalize effectively across diverse educational contexts.

Notes for Practice

- Predictive modeling in learning analytics often struggles with generalizability and scalability due to variations in context, data availability, and implementation across institutions.
- This paper introduces modelability as a design strategy for creating learning ecosystems that enable the development of predictive models that are both generalizable and scalable, demonstrated through a cross-institutional case study using CourseKata.
- The findings suggest that intentionally designed, instrumented learning platforms can support the development of early interventions by enabling predictive models that transfer effectively across diverse educational contexts.

Keywords: Modelability, generalizability, scalability, cross-institutional research, predictive learning analytics

Submitted: 04/06/2025 — **Accepted:** 30/12/2025 — **Published:** 14/03/2026

Corresponding author [1]Email: alicex@g.ucla.edu Address: Department of Psychology, University of California, Los Angeles (UCLA), 1285 Franz Hall Box 951563, Los Angeles, CA 90095, USA. ORCID iD: https://orcid.org/0000-0001-8111-0700
[2]Email: icy.zhang@wisc.edu Address: Department of Educational Psychology, University of Wisconsin–Madison, 1025 W Johnson St, Madison, WI 53706, USA. ORCID iD: https://orcid.org/0000-0003-3423-6794
[3]Email: adam@coursekata.org Address: Fund for the City of New York, 121 Avenue of the Americas, 6th F1, New York, NY 10013, USA. ORCID iD: https://orcid.org/0000-0001-7881-8652
[4]Email: stigler@ucla.edu Address: Department of Psychology, University of California, Los Angeles (UCLA), 1285 Franz Hall Box 951563, Los Angeles, CA 90095, USA. ORCID iD: https://orcid.org/0000-0001-6107-7827

1. Introduction

Online learning has transformed education by making learning more accessible, leading to growing adoption, and offering new ways to understand the learning process. Meanwhile, it also generates vast amounts of learner data that can be captured, interpreted, and effectively used to enhance educational practices. In response, the field of learning analytics (LA) has emerged, focusing on "*the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs*" (Conole et al., 2011).

Despite its promise, the field of learning analytics struggles with generalizability and scalability (Mathrani et al., 2021; Moreno-Marcos, 2018). On the one hand, variation in curricula, student populations, and instructional strategies makes it nearly impossible to develop broadly applicable models. On the other hand, overly prescriptive solutions that work in a single context may demand uniformity, potentially alienating educators who need flexibility to adapt instruction to their students' needs. This tension—between models that are too general and those that are too specific—impedes the creation and

implementation of scalable solutions that function effectively across diverse learning contexts, reducing the long-term impact of learning analytics on education.

In this paper, we begin by discussing why the issue of generalizability persists and remains difficult to resolve, and how it is closely intertwined with scalability. We then propose a solution centered on designing an ecological research context that fosters collaboration among multiple stakeholders, bridges research and practice, and promotes generalizable and scalable solutions to enhance student learning. Next, we introduce an implementation of the proposed approach to demonstrate its practical application. Finally, we present a case study to inspire learning analytics researchers as to how they can leverage this framework in their own work.

1.1. The Generalizability Issue in Learning Analytics

Learning analytics has the potential to improve education by using data to gain insight into how students learn. A central part of learning analytics is predictive modeling, in which factors such as demographic information, prior academic performance, and behaviors in the current class are used to predict learning outcomes (e.g., course grades, or identification of at-risk students. For a review, see Sghir et al., 2023). Once a model is developed, it can then be applied to future data to generate predictions to inform decision-making. However, since model development relies on historical data while its real-world application involves future, unseen data, discrepancies between the two can compromise model performance (Storkey, 2009).

In theory, if the future data closely resembled the data the model was developed on in distribution and structure, the model's generalizability should be high. However, in reality, learning is dynamic and varies across contexts, making it difficult for a model developed in one context to generalize well to another (Ge et al., 2014). For example, across institutions, differences in teaching and student characteristics can affect model generalizability. Even within the same institution and the same course, different instructors may vary in how they assign coursework and structure the class, leading to differences in student learning behaviors (Suleman et al., 2022; Trautwein et al., 2009).

Measures that appear similar on the surface may not have the same meaning across contexts. For example, in a course where online discussions are required and encouraged, the number of times a student uses the discussion board may be a meaningful and predictive feature of course outcome (e.g., Kim et al., 2016). However, in a course where online discussion is optional or unavailable, this metric would not hold the same predictive value. Students who participate voluntarily may have different motivations than those required to do so (Sharma et al., 2016). In this case, the same behavior can stem from different underlying reasons and carry different implications depending on the learning context.

The field of learning analytics has recognized the impact of contextual variations, and some researchers argue that a "one-size-fits-all" approach should not be promoted (Gašević et al., 2016). In practice, most learning analytics studies fall on the "specific" end of the "specific–general" spectrum of generalizability, developing models based on one or a few courses within a particular discipline, often within a single institution. Some may believe this is the best we can do. Although such models may not generalize, they still might be very useful in their specific contexts. And besides, the lack of access to diverse, multi-institutional datasets makes it hard to develop truly generalizable models.

In our view, however, it would be a mistake to give up the quest for generalizability, at least not without a fight! Working in limited contexts with only minor variation not only produces models that do not generalize but also makes it more difficult to develop useful theories of teaching and learning. Without data from multiple and varied contexts, researchers will struggle to identify broader, transferable mechanisms for improving learning, which in turn will limit the ability to develop cost-effective, widely applicable interventions (Tsai et al., 2020).

In addition, while well-resourced institutions have the necessary policies, infrastructure, and funding to support their own learning analytics research, others with fewer resources may struggle to do so. This disparity contradicts the broader goal of online learning: to ensure high-quality education and equitable opportunities for all learners (Farley & Burbules, 2022; Tate & Warschauer, 2022). Instead of democratizing education, current learning analytics practices risk reinforcing existing inequalities, as only a limited number of institutions are positioned to benefit from these advancements (Ifenthaler, 2016).

1.2. Generalizability and Scalability

Besides generalizability, another persistent and closely-related issue in learning analytics research is scalability. Whereas generalizability refers to "the ability to generalize the results to different settings, groups, and situations" (Munir et al., 2014), scalability refers to the ability of a learning analytics solution to efficiently handle increasing amounts of data without compromising performance (Pelánek et al., 2020). Scalability involves both the computational efficiency required to process larger datasets or more complex analyses and the capacity to expand the development and application of the solution across diverse settings. Scalability goes a step further than generalizability by addressing how research can scale from lab studies to real-world adoption; it's about growth, and applying research findings in practical, large-scale contexts.

Generalizability and scalability are closely related. A system that generalizes well is easier to scale, as it can be reliably applied to new contexts. Conversely, without scalability, it will not be possible to develop and test the generalizability of models across diverse, real-world settings. Both generalizability and scalability of results are crucial for learning analytics

research, where many researchers play dual roles: as learning scientists, seeking to understand the mechanisms behind human learning; and as educators, striving to translate knowledge about learning into meaningful practices that support learners. Yet, despite acknowledgements of this fact, most learning analytics research projects are still small-scale, exploratory studies (Dawson et al., 2019) that fail to capture the complexity of education or deliver meaningful impact.

This disconnect is illustrated by a review from Dawson and colleagues (2019), who analyzed over 500 papers published in the Learning Analytics and Knowledge (LAK) conference proceedings and the Journal of Learning Analytics (JLA) from 2011 to 2018. They found the vast majority of studies were conducted at the course and project level, while only a minority extended to the institutional level or beyond. Many studies are also exploratory in nature, typically focusing on developing specific models based on particular datasets to predict student outcomes. This contrasts with evaluative research, which seeks to assess the effectiveness of implemented models or interventions in real-world settings.

It's unsurprising that the small-scale and exploratory nature of many learning analytics studies often go hand in hand. When researchers have access only to limited, highly specific datasets—often confined to a single course or institution—it becomes difficult to pursue evaluations beyond that narrow context. As a result, many studies focus on exploring patterns and developing models rather than validating generalizable outcomes. While overfitting may be one factor contributing to a lack of generalization, it is not the primary cause. It has long been a well-studied issue with established techniques to address it (Ying, 2019). A more persistent challenge stems from the inherent limitations of small-scale studies that fail to account for the diversity of real-world settings.

Establishing and rigorously evaluating generalizability requires the involvement of multiple stakeholders, something that small-scale, exploratory studies, and the field of learning analytics in general, often lack (Samuelsen et al.,2019). For instance, implementing a new feature in a learning management system (LMS) may require approval from instructors and technical support from developers. While researchers may aim for large-scale evaluations, those working in isolation often lack access to critical resources needed to implement their models beyond their immediate settings.

2. Building a World More Suitable to Being Modeled

Improving generalizability is not a new topic in learning analytics research, and significant efforts have been made on the algorithmic side. Traditional models (e.g., linear regression), assume static, linear relationships, which may not accurately reflect the dynamic and nonlinear nature of learning processes. More advanced approaches, such as time-series models (Brooks et al., 2015), can better capture the temporal aspects of learning, making predictions more adaptable to students' learning progress. Additionally, efforts have been made to prevent model overfitting to specific contexts through data preprocessing and sampling techniques, helping to mitigate biases toward particular subgroups (Maharana et al., 2022).

However, these algorithmic improvements have inherent limitations and can only enhance generalizability within the constraints of the data itself. The well-known metaphor "Garbage In, Garbage Out" (GIGO), coined in the late 1950s by George Fuechsel, an IBM programmer, states that in any system the quality of output is determined by the quality of the input (Lidwell et al., 2010). This reminds us that a model is only as good as the data it is trained on. In the context of learning analytics, even the most advanced models cannot generalize effectively if their training data is highly context dependent. Moreover, advanced algorithms often come with added complexity and are harder to interpret (Lipton, 2018), which can raise accountability and transparency concerns, thereby hindering their adoption (Shin & Park, 2019).

An alternative solution to the generalizability challenge is to shift the focus from improving the modeling techniques to improving the affordances of the world to be more amenable to being modeled. A more modelable world is one that is designed with modeling and generalizability in mind. For example, educational platforms can be purposefully designed to produce comparable and interoperable data across diverse settings. As more comparable data is collected from varied contexts, the resulting models can better account for local variation and generalize more effectively across different learners and learning contexts.

As Ronald Aylmer Fisher (1938) famously put it, "To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem." This highlights the importance of structuring the conditions under which data is generated from the start rather than relying on complex statistical techniques after the fact. By addressing the generalizability issue at its source, we can enhance generalizability *before* relying on post hoc algorithmic improvement to handle the limitations arising from highly contextualized data.

2.1. What Makes a World More Modelable

We define *modelability* as the extent to which a system is designed to support the development of models that are generalizable. We argue that a highly modelable system typically includes the following three features: 1) valid and interpretable measurements, 2) scalable and stable implementation, and 3) a collaborative research–practice–technology ecosystem, as illustrated in Figure 1.

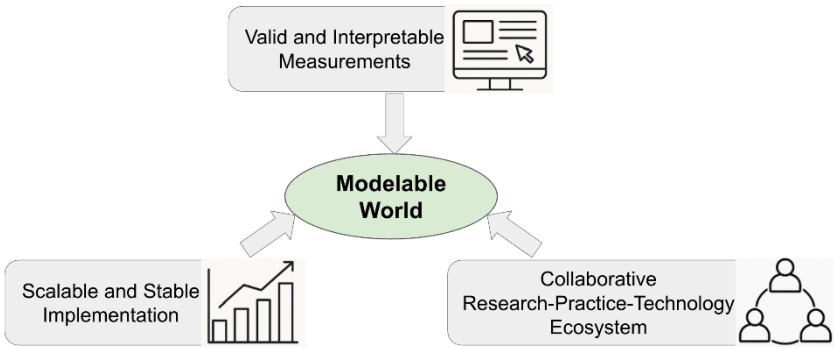

Figure 1. Key Design Principles for Building a Modelable World.

2.1.1. Valid and Interpretable Measurements

Learning analytics aims to identify factors that contribute to variability in learning events, quantify the influence of those factors, and recommend changes in learning practices that are beneficial outcomes for learners (Winne, 2020). Achieving these goals, however, depends on a critical prerequisite: the availability of valid and interpretable measurements.

Valid measurements require empirical evidence and theoretical rationale to justify the conclusions and decisions made from them (Messick, 1989). In the context of learning analytics, trace data that capture learners' interactions with digital learning environments (e.g., time spent, actions taken, tools used) have emerged as a promising source of real-time, unobtrusive evidence. They help overcome the limitations of reflective measures, which are often prone to inaccurate recall and bias (Winne, 2010). However, logging behavior alone is not sufficient, as it only captures what learners do but not necessarily the cognitive or metacognitive processes underlying those actions.

To strengthen the validity of inferences drawn from trace data, it is important to triangulate with other data sources that offer more direct insight into learners' thinking (Fan et al., 2022). This approach aligns with principles of improvement science, which emphasize using data not only to evaluate outcomes but also to inform iterative refinement of systems in authentic settings (Langley et al., 2009). Implementing this framework requires three types of measures (Lewis, 2015): 1) Outcome measures, which typically focus on learning indicators such as performance, completion, or concept mastery; 2) Process measures, which may include behavioral indicators that reflect how students engage with learning activities, for example, time spent on engaging with online textbooks; and 3) Balancing measures, which are essential to ensure that efforts to improve one aspect of learning do not inadvertently harm others, such as diminishing students' future interest in the subject.

This framework provides a useful lens for guiding the design of modelable systems for learning analytics, which must be capable of generating all three types of measures. For example, such systems should include trace data that capture detailed learner behaviors, along with additional data sources that offer complementary insights into cognitive and metacognitive processes, such as embedded assessments and self-report measures. When these sources are well-aligned, they can collectively enhance the validity of inferences and, ultimately, support the development of learning theory.

2.1.2. Scalable and Stable Implementation

While valid and interpretable measurements are important, they only become truly useful for building generalizable models when implemented at scale and with consistency. Scalable systems allow researchers to collect data across diverse learners and learning environments, enabling the identification of generalizable relationships between measures and learning outcomes.

Scalable implementation requires thoughtful decisions about which aspects of the learning environment should remain constant and which should vary, based on the research goal, before implementing the system across multiple contexts. For example, in an early-warning model that identifies at-risk students, holding the curriculum and assessment structure constant allows observed learning behaviors to more accurately reflect individual differences rather than noise introduced by varying course designs. At the same time, collecting standardized behavioral data across a wide range of contexts provides the model with the diversity it needs to generalize and scale effectively.

Conversely, if the goal is to build a predictive model that recommends effective learning materials for specific student profiles, it can be more useful to hold the student population relatively constant while systematically varying the instructional content. Controlling for learner characteristics ensures that differences in learning outcomes can be more confidently attributed to the instructional materials themselves, allowing the model to identify which content features are most predictive of success for a given group of students.

In addition to scaling across contexts, implementation should also remain stable over a sustained period. If key elements are modified too frequently, it becomes difficult to validate findings longitudinally. For example, during the coronavirus disease 2019 (COVID-19) pandemic, students' lives were significantly disrupted, and their learning behaviors changed

accordingly (Hu, 2022). Such external factors may reveal temporary relationships between learning measures and outcomes, but these patterns are not necessarily generalizable over time and should be further tested under more stable conditions.

In sum, a modelable system is not just about reaching more learners or institutions; it requires designing for consistency in key elements while allowing purposeful variation in others, to generate the structured, interpretable data needed to develop robust, generalizable models of learning.

2.1.3. Collaborative Research-Practice-Technology Ecosystem

Valid and interpretable measurement, along with scalable and stable implementation, are necessary, but not sufficient, conditions for building a truly modelable system. These elements cannot be sustained by simply handing a platform to instructors and expecting widespread adoption. Just like leaving flyers on a doorstep, people won't necessarily read them. A platform with high quality instructional content may be available, but instructors might choose not to use it. And even when the platform is adopted, there is still no guarantee it will be used in a way that produces accessible, interpretable data. Furthermore, even when data are collected and made available to researchers, model development alone does not ensure improvements in student learning, especially without the contextual knowledge and sustained engagement of practitioners.

Creating a modelable ecosystem requires coordinated effort across the research–practice–technology partnerships (RPTPs). This framework builds on the concept of research–practice partnerships (RPPs, Coburn & Penuel, 2016) which emphasize sustained collaboration between researchers and practitioners to improve educational practice, while also recognizing the critical role of technological infrastructure in modern learning systems (Gudanescu, 2010). In RPTPs, institutions play a critical role in supporting implementation at scale, facilitating access to student populations, and enabling long-term data collection. Instructors provide contextual insight into classroom practices, helping researchers interpret measures more meaningfully. Researchers bring theoretical frameworks and methodological tools to guide model design. Developers, in turn, contribute the technical infrastructure needed to translate insights into usable tools or interventions.

Only through such interdisciplinary collaboration can learning analytics move beyond isolated studies and become an engine for sustainable, large-scale educational improvement.

3. CourseKata: A Design for Producing More Modelable World

The challenges of generalizability and scalability in learning analytics reflect broader issues in social science and education research. Insights into learning are often difficult to apply in practice, slow to scale, or disconnected from the systems that could support their use—limiting their real-world impact. In response to these challenges, Stigler et al. (2020) proposed building a research and development (R&D) ecosystem around the continuous improvement of educational materials and their implementation—a model they called the *Better Book* approach.

An example of this approach is CourseKata, a non-profit education R&D company founded in 2017 that embodies the Better Book philosophy and has evolved into a system well-aligned with the principles of modelability. CourseKata's mission is to bring together researchers, designers, and educators to work on bringing transferable learning to scale. CourseKata's first product is a set of interactive, online textbooks for introductory statistics, which are implemented in both colleges and high schools. Starting with Version 1, CourseKata applies improvement science methodologies and a collaborative approach to the continuous incremental improvement of the books over time.

This continuous improvement process is supported by the CourseKata technology platform, developed in parallel since the organization's founding, which integrates content authoring, delivery, data management, and research; all of which collectively enable continuous improvement of students' learning. Content (the online textbook) is delivered to students through an LMS, where teachers have access to identified student data. De-identified data are stored on CourseKata's server and can be accessed by researchers. As a founding partner in the U.S. National Science Foundation's SafeInsights research hub, CourseKata ensures that data is shared under rigorous privacy standards that protect student information.

CourseKata actively fosters collaborative environments where researchers, educators, and developers work together, each contributing their unique expertise to advance learning analytics research and practice. With support from teachers and designers, researchers collect data from diverse educational contexts, conduct random-assignment experiments within real classes, test models' generalizability across different settings, and implement effective solutions at scale, while simultaneously creating a continuous improvement loop for refining CourseKata's content.

3.1. CourseKata's Approach to Support Modelability

CourseKata exemplifies a modelable system, supported by collaboration among researchers, practitioners, and developers, with specific affordances that make it amenable to the development of generalizable predictive models at scale. Specifically, it (1) builds a large quantity of standardized measures into the online textbook, and makes that data easily available outside of institutional LMSs; (2) strikes a balance between consistency and variability by holding online textbook content constant to facilitate comparisons across contexts while allowing for variation in implementation; and (3) as it scales, it increasingly provides data across a large and diverse range of contexts.

3.1.1. Building Standardized Measures into the Online Textbook

Many learning analytics researchers rely on LMS data to develop their models (e.g., Conijn et al., 2016; Firat, 2016; Hernández-García et al., 2024; Mwalumbwe & Mtebe, 2017); however, different LMS platforms capture different measures, and there is no easy way to aggregate and compare data across systems.

As a result, many studies use distinct features and are limited to a single institution, restricting generalizability. CourseKata has met this challenge by designing a data collection process that captures deidentified and standardized student interaction and learning data in a data warehouse that exists outside of the various LMSs in which CourseKata is delivered. These data have the potential to support valid and interpretable measurements.

The CourseKata textbook is fully instrumented to yield data relevant to students' interactions, learning, and thinking as they progress through the course. Students' interactions with the textbook are monitored in three ways. Trace data is collected using enGauge (Blake & Stigler, 2021), a tool built into the platform that tracks user activities (e.g., scrolling, cursor movement, and keyboard activity) in real time. When a user first accesses a page, they are labeled as *engaged*, and a timestamp is recorded. If no activity is detected for more than two minutes, a label of *idle* is generated, and the time is recorded. Once activity resumes, a new label of engaged is generated. If the user clicks out of the page and enters a new tab or application, they are labeled *off-page,* and the time is recorded. Similarly, if they close the book page entirely, a *close-page* is generated. These timestamps provide a detailed view of how students engage with the online textbook.

A second source of data comes from students' responses to the more than 1,500 embedded formative assessments and coding activities that are included in the textbook. These questions are designed to support the development of transferable knowledge and foster a coherent understanding of the subject through "practicing connections" (Fries et al., 2020). The questions take various forms such as multiple-choice, matching, and ordering, all of which are auto-scored, as well as open-response questions. Formative assessments, which are conducted during the learning process, have been shown to be strong predictors of learning outcomes (Bulut et al., 2023). By combining detailed response data with reading patterns, we can create a comprehensive overview of when and how students engage in learning throughout the course.

A third source of data comes from embedded self-report surveys administered at multiple points throughout the course. These surveys capture changes in key psychological constructs such as student motivation and future interest in the domain (Sutter et al., 2024).

Trace data, embedded formative assessments, and survey responses can generate process measures that capture students' learning trajectories such as patterns of engagement, attempts, and strategy use, offering rich information about cognitive and metacognitive processes. The role of each data source may vary depending on the specific research goal. For example, formative assessments can serve as outcome measures to assess mastery of specific learning objectives, while survey responses can be used to generate balancing measures such as detecting unintended declining interest. Together, these data sources support the development of valid, interpretable, and theory-informed measures and models.

3.1.2. Keeping the Textbook Constant to Facilitate Comparisons Across Contexts

Learning is often domain-specific (Tricot & Sweller, 2014), meaning that a student who excels in one area may not necessarily perform well in another. This variability makes it difficult to develop a universally generalizable model of learning across different domains. Even within the same subject, differences across curricula and textbooks can pose challenges to generalizability. For instance, CourseKata teaches statistics using R programming (Tucker et al., 2023), while another curriculum may teach the same subject without coding. This disparity not only leads to differences in how students engage with and learn the material but also affects how we define "learning," with varying learning objectives established at the outset.

To address these challenges, CourseKata adopts a common online textbook that aligns learning objectives across students and institutions, enabling more meaningful comparisons and predictions. By holding the instructional content constant, we can reduce curriculum-related variability, allowing differences in learning outcomes to be more confidently attributed to individual characteristics and other contextual factors beyond the curriculum.

Importantly, keeping the textbook constant also creates a foundation for generating derivative versions of the original content, designed to test specific instructional improvements. In this modelable system, the goal is not just to model learning outcomes, but to support the iterative refinement of both the models and the instructional materials themselves. A shared, stable textbook allows researchers to introduce, test, and identify changes that are most effective in improving student learning. In this way, the system becomes capable of learning and evolving over time, based on evidence generated through its own use.

3.1.3. Scaling the Textbook Across a Large and Diverse Number of Contexts

A priority at CourseKata has been to spread the use of its materials across many types of institutions (e.g., research universities, liberal arts colleges, and community colleges where both *learners* and *learning environments* vary to a large extent), as well as the many different disciplines where statistics is taught (e.g., mathematics, statistics, psychology, political science, and business). Expanding the range of settings where the book is implemented is one important step toward developing a model that is less biased and more beneficial to a broader range of learners (Wilson et al., 2017).

Supporting implementation across a large variety of contexts has resulted in a number of design decisions on the part of CourseKata. Notably, CourseKata integrates easily into the LMS platforms that teachers are already familiar with, which accelerates adoption and leads to larger, more diverse samples, providing a stronger foundation for building generalizable learning models and scalability. It has proven important to study the challenges teachers face implementing the materials and then to design improvements to reduce such challenges.

CourseKata has also created professional development (PD) programs that equip teachers with the necessary skills to teach the course. These PD programs foster a community where teachers can discuss questions, address concerns, and support one another in delivering the curriculum. CourseKata actively works to build a strong social network among teachers as well as between teachers and researchers, fostering a sense of belonging within the community.

4. Current Study: A Case Study of How Predictive Models Are Developed and Evaluated in the CourseKata Ecosystem

With the increased adoption of CourseKata in higher education and K–12, we began exploring how to leverage it for conducting generalizable and scalable learning analytics research. In the rest of this paper, we present a case study in which we explore and illustrate the affordances of a community such as CourseKata for developing an early prediction model of students' final course grades. We developed predictive models using data collected on CourseKata and then evaluated the cross-institutional generalizability of the models, testing model performance in an institution different from the one in which the model was developed. We hope the case study reported here sparks interest among learning analytics researchers who struggle with establishing and evaluating generalizability and scaling their findings.

Specifically, we developed a model detecting at-risk students early in the courses, using behavioral-only metrics derived from students' uses of an online textbook in an introductory-level statistics course delivered in a flipped classroom setting. We then evaluated the model's generalizability in a different course using the same textbook taught by another instructor, at a different institution, and over a different course length. For this case study, we focus on the following research questions:

RQ1: What is the predictive performance of models using behavioral-only metrics, engineered from students' textbook interactions, to predict final course grades early in the term in a flipped classroom setting?

RQ2: How well do such models generalize across different contexts (i.e., two statistics courses using the same online textbook, taught at different institutions)?

5. Method

5.1. Data Collection

Data were collected from two introductory statistics courses at two different public research universities in California. Both courses used the same CourseKata *Statistics and Data Science* online textbook (Son & Stigler, 2017–2025), delivered through Canvas. Both courses followed a flipped classroom model, where students completed interactive textbook readings of the corresponding chapters prior to attending live lectures and discussion sessions.

Course 1 (Spring 2022) ran for ten weeks. 183 students completed the course and received a final grade. Course 2 (Summer 2022) was an accelerated six-week session with 60 students who completed the course and received a final grade. Course 1 included a weekly discussion section in addition to lectures, whereas Course 2 did not. Table 1 summarizes the demographic composition of students in the two courses.

Table 1. Student Composition of the Two Courses

	Course 1	Course 2
Number of students	183	60
Gender identity	66% female, 29% male, 5% others or not reported	66% female, 28% male, 6% others or not reported
Age	Mean age 19.4 years (SD = 1.6)	Mean age 20.5 years (SD = 3.2)
Ethnicity	16% White, 46% Asian or Asian American, 26% Hispanic, Latino, or Spanish Origin, 4% Black or African American, 7% Others	13% White, 48% Asian or Asian American, 30% Hispanic, Latino, or Spanish Origin, 3% Black or African American, 5% Others
Class standing	57% freshman, 19% sophomore, 14% junior, 9% senior, 1% other	3% freshman, 31% sophomore, 52% junior, 10% senior, 3% Other
Transfer students	10%	14%
No programming experience	56%	31%

Time-stamped events of *engaged*, *idle, off-page,* and *close-page* are processed to construct reading *sessions*, which will be described in detail in the reading-related features section. These sessions provide a standardized measure of how students spent their time engaging with the online textbook throughout the course.

Formative assessment performance data were generated from students' responses to the embedded questions including multiple-choice, matching, ordering, and open-response formats, as well as coding exercises within the online textbook. These assessments appeared both within lesson pages and as quizzes at the end of select chapters. Student performance on non-coding questions was auto-scored on their first attempt, with correct answers scored as 1 and incorrect answers as 0, except for the open-response questions, which were not scored. Although students could reset a page and attempt the questions again, only their first attempt was recorded, as the correct answers were revealed immediately afterward.

For the coding exercises, students were allowed to edit and submit their code multiple times. Each submission, along with whether it was correct or not, was recorded. Students received immediate feedback on whether their submission was correct, but the correct answer was not provided if their submission was incorrect. Because of this, all coding attempts were scored and recorded, which contrasts with the non-coding questions for which only the first attempt was recorded, as the correct answers were revealed after that initial attempt. These formative assessments provide insight into how students engaged with and initially processed course content during their first encounter with the material. In both courses, performance on formative assessments did *not* directly contribute to students' final grades. Only completion counted toward a small portion of the overall grade, and nearly all students received full credit for this component.

5.2. Data Labeling

Students' final letter grades, provided by the instructors as supplementary materials to CourseKata, were used as the outcome variable for the prediction models. One of the two courses provided only final letter grades (without numeric scores), which are inherently ordinal and do not represent equal intervals. Given this constraint, we used a classification approach rather than regression to ensure consistency across both datasets and to more appropriately model the structure of the outcome variable.

As this is an introductory course designed to build foundational skills, students earning below an A- are flagged for additional support, and those below a B- warranting the greatest attention. In practice, interventions are often targeted by grade bands rather than exact scores, so even if performance were predicted on a continuous scale, it would ultimately need to be converted into categories for decision-making.

Based on these considerations, we collapsed these grades into three performance categories. Class 0 (high achievers): A-, A, or A+; Class 1 (moderate performers): B-, B, or B+; Class 2 (at-risk students): C+ or lower. In Course 1, 117 students were assigned to Class 0, 24 to Class 1, and 42 to Class 2. In Course 2, the counts were 41, 10, and 9, respectively.

5.3. Feature Generation

Data were prepared and analyzed using the scikit-learn package (Pedregosa et al., 2011) in Python 3 (Van Rossum & Drake, 2009).

5.3.1. Feature Selection

To model student learning over the duration of the course and predict their final grade category, we selected predictor variables that were available at the chapter level. This level of granularity allowed us to track learning as it unfolded and supported the goal of making early predictions. Because students interacted with the textbook as the course progressed, chapter-level features—such as early reading behavior—became available in near real time, enabling timely identification of at-risk students. Specifically, we constructed 11 chapter-level features: six related to reading behavior and five related to formative assessment performance.

Reading-Related Features. Time-related learning behaviors are critical in online learning environments, where students have considerable autonomy in deciding when and how they engage with material. Prior research has shown the importance of time management for academic success (Mega et al., 2014; Kim et al., 2018; van Halem et al., 2020; Zimmerman, 2012). Yet the amount of time spent alone may not provide the full picture; how that time is used also plays a key role. For instance, even when students devote similar amounts of time to studying, their performance can differ depending on how they structure their study sessions and maintain focus throughout each one.

To capture more nuanced dimensions of time use, each learner's interactions with the book were segmented into a series of reading sessions based on timestamped activity data. Sessions served as the basic unit of analysis for describing how students structured their learning over time. Compared to raw timestamp data, session-based features offer a more standardized and interpretable representation of student behavior, making it easier to analyze and compare learning patterns across individuals. Previous studies have shown that dividing students' textbook engagement into sessions enhances our understanding of learning behaviors in online settings (Kovanović et al., 2015; Liu et al., 2015; Maldonado-Mahauad et al., 2018). Moreover, session-related features—such as the frequency and duration—have been shown to predict academic performance (Junco & Clem, 2015; Xu et al., 2023).

A session begins at the initial *engaged* timestamp and continues until the next timestamp after which no activity is detected for more than 10 minutes. If a user resumes their activity within 10 minutes, it is still considered part of the same session. Thus, a session may include periods of active engagement, brief idleness, short off-page intervals, or transitions between pages, as long as no continuous period of inactivity exceeds 10 minutes. Once a session ends, the next *engaged* timestamp marks the beginning of a new session. Within a single session, the learner could spend time on one or multiple pages. See Figure 2 for an illustration of how a session is defined.

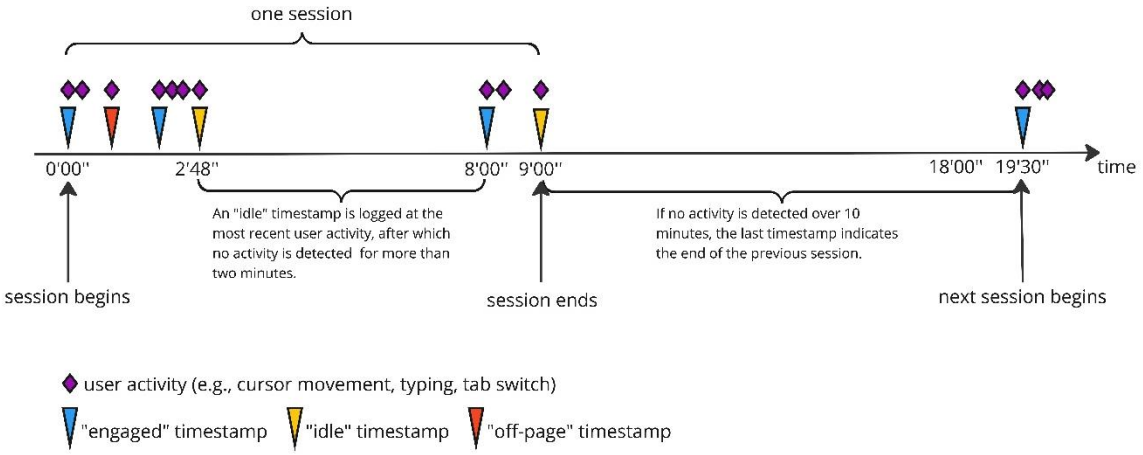

Figure 2. Illustration of How a Session is Defined. Note that all user activities (e.g., cursor movement, typing, tab switching—labeled as purple diamonds on the graph) are detected by enGauge, but timestamps (e.g., engaged, idle, off-page—labeled as triangles) are only recorded when the user's status changes.

We removed any session shorter than one minute, as meaningful learning was unlikely to occur within such a brief period. These very short sessions also rarely included interactions with embedded questions. Additionally, sessions that took place within end-of-chapter quizzes were excluded from the generation of reading-related features. Although students may read during these quizzes, their engagement is often driven by the immediate goal of answering questions rather than by reflective reading. To better capture natural reading behavior, we focused only on sessions occurring on regular textbook pages.

After organizing the reading time into sessions, we extracted six reading-related features for each chapter:

1. *Session count (sess_count)*: The number of reading sessions recorded for a chapter. This reflects how students distribute their reading—more sessions may suggest deliberate spacing or difficulty sustaining attention, while fewer sessions may indicate cramming or longer, focused study periods.

2. *Average session duration (sess_avg_dur)*: The average length of all sessions for a given chapter, measured in minutes. Longer sessions might indicate sustained attention or slow processing, while shorter ones may reflect efficient studying or brief, fragmented engagement.

3. *Maximum session duration (sess_max_dur)*: The duration of the longest session for a chapter, measured in minutes. This provides insight into a student's capacity for sustained engagement. A long sess_max_dur might reflect deep focus or last-minute cramming.

4. *Pages per session (pages_per_sess)*: The mean number of textbook pages visited per session within a chapter. This estimates the amount of content covered per session. Higher values may suggest fast reading or skimming, while lower values may indicate slower, more thorough reading.

5. *Total reading time (read_time_total)*: The sum of all session durations for a chapter, in minutes. This captures the overall time spent reading. While greater total time may suggest effort and engagement, it may also reflect inefficient study strategies.

6. *Engagement ratio (engage_ratio)*: The proportion of time a student was actively engaged during reading sessions, calculated as total engaged time divided by total reading time (including idle or off-page time under 15 minutes). This indicates how focused students were during reading. Higher values indicate sustained attention, while lower values may reflect distraction or multitasking. See Figure 3 for an illustration of each component in the calculation of the engagement ratio.

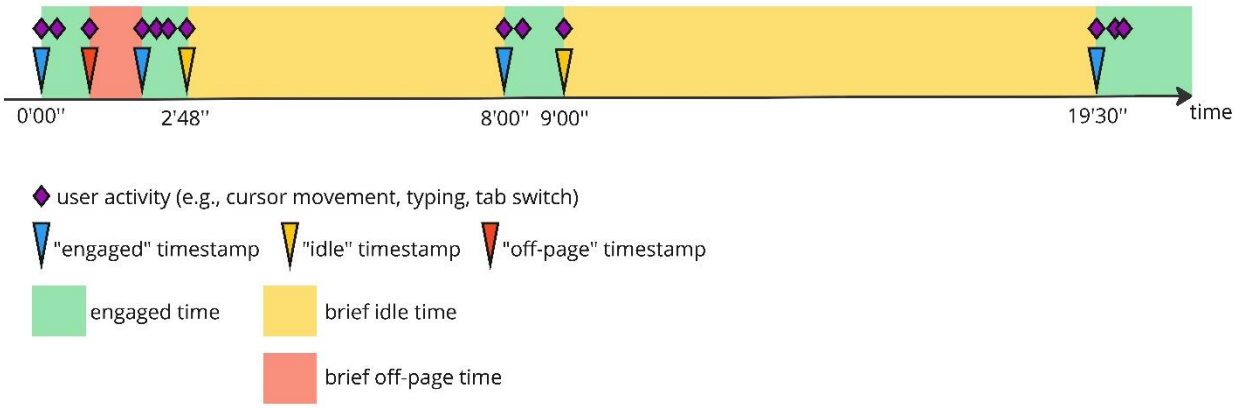

Figure 3. Illustration of the Engagement Ratio (engage_ratio). Note that the ratio of engaged time to total reading time (er) for each chapter is calculated as the sum of engaged time divided by the total of engaged time (highlighted in green), idle time (highlighted in yellow), and off-page time (highlighted in red).

Formative Assessment-Related Features. In addition to the reading-related features, we also computed five formative assessment–related measures. Some of these are based on students' responses to questions embedded within the regular textbook pages, while others capture performance on end-of-chapter quizzes, together representing students' learning progress at different stages of the learning process.

1. *Quiz score (quiz_score)*: The total score from the first end-of-chapter quiz, which is available only for Chapters 1 through 8, as the textbook does not include end-of-chapter quizzes for Chapters 9 and beyond. For chapters with two quizzes, only the first was used, as the second was highly similar in assessed content. This score reflects how well students consolidated and applied the material after reading.

2. *First attempt accuracy (first_attempt_acc)*: The percentage of embedded questions and coding exercises answered correctly on the first attempt. This measure captures initial understanding and may indicate preparation, attentiveness, or ease with the material.

3. *Final attempt accuracy (final_attempt_acc)*: The percentage of embedded questions and coding exercises answered correctly on the final attempt. For coding exercises, where students could revise and resubmit without seeing the correct answer, this score may reflect persistence (DiNapoli & Miller, 2022). For embedded questions, where correct answers were shown after the first attempt, final_attempt_acc is equal to first_attempt_acc.

4. *Attempts per question (avg_attempts)*: The average number of submissions per question, particularly relevant for coding exercises where students could submit multiple times. Higher values may reflect perseverance or difficulty with the material.

5. *Summary length (summary_word_count)*: The number of words in students' written chapter summaries, available for the first nine chapters. Longer summaries may reflect greater effort or more detailed recall of the material, while shorter summaries could indicate more efficient synthesis—or, in some cases, lower motivation to complete the task.

For a summary of all features, see Table 2. In total, the model includes 132 features, with 11 features for each of the 12 chapters. For formative assessment-related features, the quiz score (quiz_score) and summary length

(summary_word_count)—available only through Chapters 8 and 9, respectively—were set to 0 for all students in the remaining chapters. This adjustment was necessary to maintain a consistent feature vector shape across all time steps in the sequence model. It is also important to note that none of the formative assessment-related measures contributed directly to the calculation of students' final grades.

Table 2. Features within Each Chapter

Metric Type	Feature (Variable)	Description
Reading-related	sess_count	Number of reading sessions
	sess_avg_dur	Average duration of all sessions
	sess_max_dur	Duration of the longest session
	pages_per_sess	Mean number of pages visited per session
	read_time_total	Sum of all session durations
	engage_ratio	Proportion of actively engaged time during reading sessions
Formative assessment-related	quiz_score	Total score on the first page of end-of-chapter review questions
	first_attempt_acc	Percentage of embedded questions and coding exercises answered correctly on the first attempt
	final_attempt_acc	Percentage of embedded questions and coding exercises answered correctly on the final attempt
	avg_attempts	Average number of submissions per question
	summary_word_count	Number of words in students' written chapter summaries

5.3.2. Train/Test Split, Oversampling, and Data Transformation

In machine learning (ML), researchers commonly use a technique called train/test split to evaluate how well a model developed with one set of data performs on a new, different set of data (Vabalas et al., 2019). This method involves randomly dividing the available data into two parts: one part is used to "train" the model (i.e., find patterns in the data that increase its predictive power), the other part to "test" the model's performance on a different set of data. Testing the model on data it hasn't seen during training helps estimate how well it is likely to perform in real-world scenarios, where the goal is to make accurate predictions on future data.

We used a variation of train/test split called *k*-fold cross-validation, a method designed to reduce the impact a particular random split could have on model evaluation (Golub et al., 1979). Rather than split the data once, we split the data into five randomly selected groups (folds), each containing 20% of the cases (in our study, each case represents one student). In each round of analysis, a model is trained on four folds (80%) and tested on the remaining fold (20%). This process is repeated five times, with each fold used as the test fold once. Results from all five rounds are averaged to estimate overall model performance.

To ensure that the training and testing sets in each round had a balanced distribution of the two outcome labels (in our case, low performers or high performers), we used a method called stratified split (Kohavi, 1995). This ensured that both the training and testing data had similar proportions of different categories, which is important because it helps the model be tested fairly on a representative mix of data.

Moreover, we used the synthetic minority over-sampling (SMOTE) technique to achieve a balanced distribution across categories in the training data (Chawla et al., 2002). Achieving this balance is important for effective model training, especially with small datasets that are prone to overfitting when they are imbalanced, which leads to poor generalization on minority class predictions (He & Garcia, 2009; Mohammed et al., 2020). This was relevant in our case, where Class 1 (moderate performers) and Class 2 (at-risk students) were underrepresented.

Finally, raw feature scores were transformed (i.e., standardized) prior to the development and testing of models. Standardization helps machine learning models converge more efficiently and improves performance, especially for models sensitive to the scale of data, e.g., neural networks (Singh & Singh, 2020). The mean and standard deviation for each feature (i.e., the standard scalers) were calculated using only the training set. *Z*-score normalization was applied to the training set before model training and finetuning, and the same scalers were subsequently used to normalize the testing set before model

evaluation. By calculating the standard scalers based solely on the training data after the train/test split, we ensured that no information from the test set "leaked" into the training process, preserving the integrity of the evaluation.

5.3.3. Data Augmentation for Early Prediction

Our training dataset included students' data across all 12 chapters. However, because we wanted to make predictions early in the course, we created an augmented dataset that would represent each student's data at the point at which they had completed each successive chapter in the book. The augmented dataset included 12 records for each student. In the first record, all data after Chapter 1 was marked as missing (nan); in the second data set, all data after Chapter 2; and so on (See Figure 4 for an illustration of data augmentation for one training record). This allowed us to build models based on data only up through the end of each chapter. While we also trained a model using data from all 12 chapters, it may rely heavily on behaviors observed in the later chapters, which cover different content and tend to be more challenging as the course progresses. However, the reading and formative assessment patterns from these later chapters are not yet available early in the course, which may limit the model's ability to make accurate early predictions (Yang et al., 2017).

Augmented data sets were created after each train/test split across the five rounds of cross-validation to prevent data leakage. If augmentation occurred before splitting, augmented traces from a single student's data could appear in both training and testing sets, allowing the model to "see" information too similar to the test set during training. This overlap would inflate accuracy by giving the model access to information it shouldn't have, compromising the integrity of the test results.

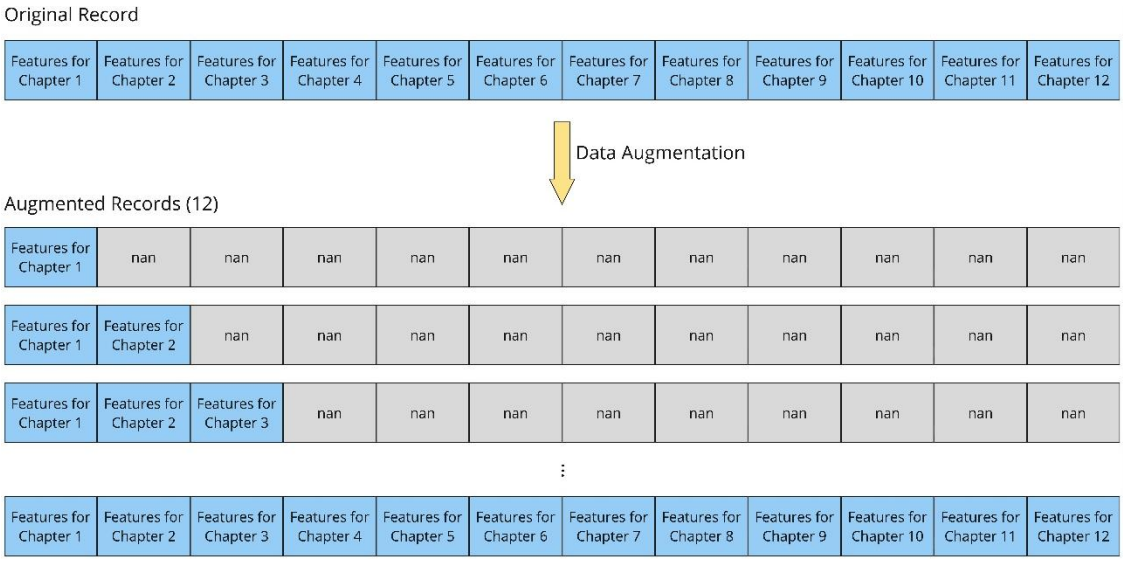

Figure 4. Illustration of Data Augmentation for One Training Record.

Most existing data augmentation approaches such as sliding windows (which create overlapping fixed-length segments) or cropping (which randomly extracts fixed-length chunks) do not reflect how student behavior naturally unfolds over time (for a review, see Alomar et al., 2023). In contrast, we generate all prefixes of each sequence (from length 1 to n), allowing the model to learn to make predictions at any point in a student's progression. This novel approach is particularly well-suited for early prediction tasks in educational settings, where decisions often need to be made before a student completes the full sequence.

5.4. Generating Early Prediction Test Datasets

Similar to the training data augmentation, twelve test datasets were created that masked data beyond each of the 12 chapters in the book. This enabled us to evaluate predictive models based on completion through each of the 12 chapters.

5.5. Building Predictive Models

After features were generated, we proceeded to train the predictive models. Models trained on full traces served as a baseline against which to assess the performance of models trained on the augmented dataset.

We created three types of models: a single Random Forest (RF) classifier, a sequence model, and a chapter ensemble model. We trained two separate models for both the single RF classifier and sequence models, one on the full traces and another on the augmented traces. The chapter ensemble model consisted of 12 separate RF classifiers; each trained on data corresponding to a specific subset of chapters. For example, the model used to predict students' final course performance based

on their trace data up through Chapter 3 used a feature vector with 33 elements (11 features multiplied by 3 chapters). The rationale for choosing each type of model will be discussed after introducing the data pipeline (Figure 5).

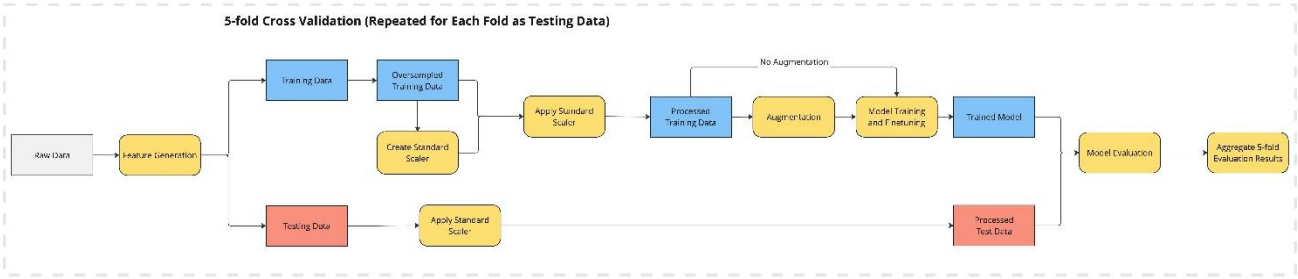

Figure 5. Data Pipeline.

As shown in the figure, after generating features from raw data, we used 5-fold cross-validation. Within each round of cross-validation, the training data were oversampled, and the features were standardized. The mean and standard deviation calculated from the training set were then used to transform the testing set to prevent "data leakage" as discussed previously. Standardized features were then used for model training and fine-tuning. In fine-tuning we used grid search, an exhaustive search over a predefined set of hyperparameters (i.e., parameters defined before training, as opposed to learned parameters that the model adjusts during the training process), to identify the optimal model configuration (i.e., best performance during fine-tuning, Bergstra & Bengio, 2012). Once the best set of hyperparameters was found, the model was trained using these hyperparameters on the training data. The model's performance was then evaluated on the testing set. This process, from standardization to model evaluation, was repeated five times in accordance with the 5-fold cross-validation.

To further reduce the impact of random splits in 5-fold cross-validation, we repeated the entire process five times, each time using a different random seed. This resulted in 25 sets of performance metrics (5 folds × 5 repetitions). The final evaluation metrics were computed by averaging the model performance across all 25 sets.

5.5.1. Random Forest Classifiers Using Original and Augmented Data

The Random Forest (RF) classifier is a popular machine learning tool that improves prediction accuracy by using a collection of decision trees rather than just one (Brooks & Thompson, 2017). A single decision tree works by repeatedly splitting data based on which features and cut points result in the best predictions based on the data. This approach can easily lead to overfitting, however, where the model fits the training data too closely and performs poorly on new data. The RF classifier reduces this overfitting by combining the predictions of multiple trees through a "majority vote", making it more reliable than a single decision tree.

Each tree in a Random Forest is unique because it is trained on a different random subset of the data—a technique known as bootstrap sampling—and, at each decision point, it considers only a random selection of features. This randomness ensures that each tree captures different patterns in the data. When the predictions of all these diverse trees are combined, the overall model becomes stronger, more balanced, and better at generalizing to new data.

In our study, we created two RF classifiers: one trained on the original dataset and another on the augmented dataset. To get the best performance from each model, we used a method of fine-tuning called grid search to test different settings, or hyperparameters, for the RF model. Specifically, we experimented with different values for the number of trees in the forest (n_estimators: 50, 100, 200) and the maximum depth of each tree (max_depth: 10, 20, no limit). Grid search involves trying various combinations of these settings to find the best possible configuration for the model, optimizing its accuracy and reliability.

5.5.2. Sequence Models Using Original Data and Augmented Data

Sequence models, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, are specialized types of machine learning models designed to work with sequential data where the order matters (Hochreiter & Schmidhuber, 1997). Examples of sequential data include sentences in a paragraph, notes in a piece of music, or, in our case, the sequence of chapters a student reads. These models are excellent at capturing patterns over sequences because they can recognize how earlier parts of the sequence influence later parts.

LSTM models tend to work best for extended sequences with long-term dependencies (Sherstinsky, 2020). However, for short sequences, like a 12-chapter textbook, RNNs perform just as well as LSTMs. And because RNNs are simpler, they are less computationally intensive, which means they require less computing power and time to train without compromising performance. For these reasons, we chose to use RNN for modeling the chapter-level sequences in the current study.

We trained two models: one on the original dataset and another on the augmented dataset. The sequence models we built consisted of three key layers (see Figure 6): 1) A mask layer to handle incomplete data by excluding "nan" positions from the

recurrent computation. For example, if a student has completed four chapters, the RNN performs four recurrent updates and ignores the remaining padded inputs, making the sequence effectively length 4. For the original dataset, this layer did not mask any values, as the sequences included features from all 12 chapters. 2) A RNN layer, from which we extracted the output of the last time step as the representation for the entire sequence. 3) A fully connected (FC) classification layer that produced the final output.

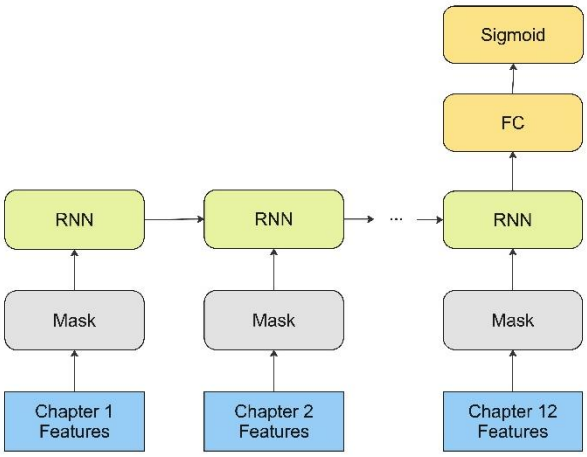

Figure 6. The Architecture of the Sequence Model.

For fine-tuning, we conducted a grid search over key hyperparameters, specifically the number of hidden units (16, 32), the type of RNN cell ("SimpleRNN" or "GRU"), and the number of units in the classification dense layer (8, 16), to optimize model performance. We trained the models using the Adam optimizer with a learning rate of 0.001 and categorical cross-entropy loss. Training was conducted with a batch size of 32 for up to 100 epochs, with early stopping applied if the validation ROC-AUC did not improve for five consecutive epochs. A dropout rate of 0.5 was applied to both the recurrent and dense layers to mitigate overfitting.

5.5.3. Chapter Ensemble Model

The chapter ensemble model is an ensemble of RF classifiers, with each classifier trained on data for each subset of successively completed chapters. Although we attempted to address the issue of incomplete traces during early prediction by using an augmented dataset, the simplicity of RF models may limit their effectiveness. RF models typically have fewer hyperparameters compared to more complex models, like neural networks. This limitation could hinder the model's ability to generalize well when dealing with the complexity introduced by varying trace lengths.

Rather than relying on more complex models, like sequence models, we can employ a divide-and-conquer approach to build an ensemble classifier (Asafuddoula et al., 2017). This method involves training multiple models, each on data with a specific sequence length, ensuring that predictions for data up to a certain chapter are made by a model trained on traces of the same length. By dividing the prediction problem in this way, each model can focus on learning patterns relevant to specific stages of student progression.

During fine-tuning, each RF model was tuned separately by adjusting the hyperparameters, specifically "n_estimators" and "max_depth."

5.6. Model Performance Evaluation

Because the outcome we want to predict has three classes, we evaluate the model's performance using a 3 x 3 confusion matrix that displays the number of correct and incorrect predictions for each class (Table 3). We used this table to generate four commonly used threshold metrics with different focuses: accuracy, precision, recall, and *F*-measure. The advantages and disadvantages of each metric are discussed in a review article by Hossin and Sulaiman (2015).

Table 3. Confusion Matrix

	Actual Class 0	Actual Class 1	Actual Class 2	Total
Predicted Class 0	True Positives 0 (TP_0)	E_{10}	E_{20}	N_{p0}
Predicted Class 1	E_{01}	True Positives 1 (TP_1)	E_{21}	N_{p1}
Predicted Class 2	E_{02}	E_{12}	True Positive 2 (TP_2)	N_{p2}
Total	N_{a0}	N_{a1}	N_{a2}	N

Accuracy (*acc*) measures the ratio of correct predictions to the total number of predictions, using the formula: $(TP_0 + TP_1 + TP_2) / N$, where N stands for the total number of cases. Precision (p_x) measures the proportion of students classified as a certain class who were actually in that class, using the formula: TP_x / N_{px}, where x can be 0, 1, or 2. Recall (r_x) measures the proportion of students in a certain class who were correctly classified as in that class, using the formula: TP_x / N_{ax}. *F*-measure ($F1_x$) is the harmonic mean of recall and precision: $2p_xr_x / (p_x + r_x)$. Unlike the arithmetic mean, which gives equal weight to both, the harmonic mean gives more weight to smaller values, effectively penalizing large imbalances between them. As a result, F1 will only be high if both precision and recall are reasonably high. If one is significantly lower, the F1 will be low. A high F1 indicates that the model is good at both correctly identifying a certain class and minimizing incorrect predictions.

To complement these metrics, we also report the area under the receiver operating characteristic curve for each class ($ROC\ AUC_x$). While F1 evaluates classification performance per class at a fixed decision threshold, ROC AUC assesses how well the model ranks examples, independent of a specific cutoff. In binary classification, a ROC curve plots the true positive rate against the false positive rate across different probability thresholds. The area under this curve (AUC) summarizes the model's ability to rank a randomly chosen positive instance higher than a randomly chosen negative one. For multi-class problems, ROC AUC is commonly extended using the One-vs-Rest (OvR) strategy (Fawcett, 2006; Provost & Domingos, 2001). This involves computing a binary ROC curve for each class by treating it as the positive class and combining all other classes as the negative class.

ROC AUC values can range from 0 to 1. A score of 0.5 indicates random guessing, while values below 0.5 suggest worse-than-random performance. AUCs between 0.5 and 0.7 imply low predictive ability, 0.7 to 0.8 indicate moderate ability, 0.8 to 0.9 reflect high ability, and scores from 0.9 to 1.0 signify excellent predictive ability (Huang et al., 2019).

While accuracy by nature provides an overall measure of model performance, the other metrics (precision, recall, F1, and AUC) are calculated per class. To report these metrics at the overall level, we report their weighted averages across classes, using the proportion of students in each class as weights.

6. Results

6.1. Model Performance within Course 1

We first calculated the performance of five models using data from Course 1: the RF classifier trained on the original dataset (full traces), the RF classifier trained on the augmented dataset, the sequence model trained on the original dataset, the sequence model trained on the augmented dataset, and the chapter ensemble model. The overall ROC AUC for each model is presented in Figure 7, left.

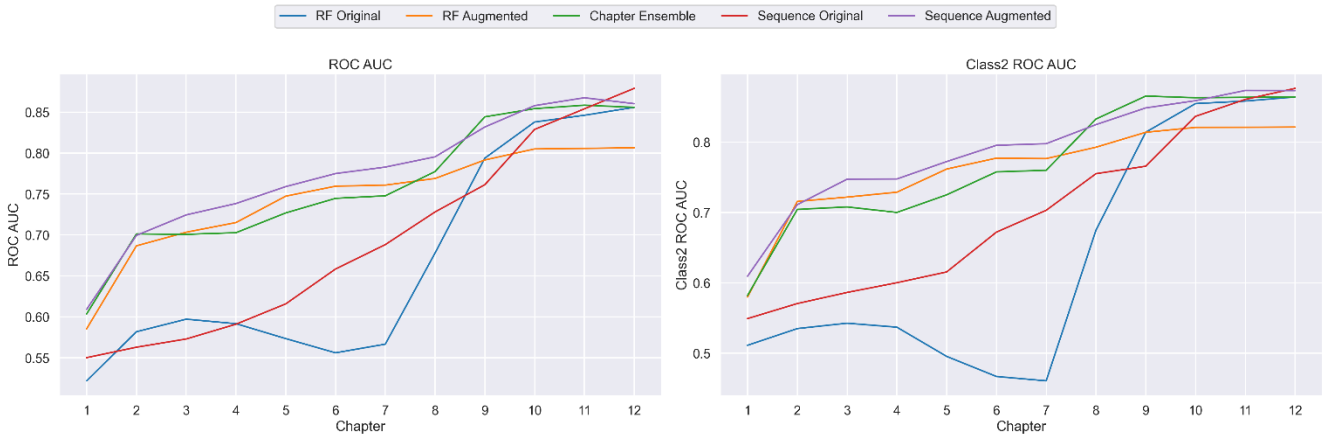

Figure 7. ROC AUC on Course 1 Data Across Five Model Variants. Note that the left graph shows the overall ROC AUC for all classes and the right graph shows the ROC AUC for Class 2 (at-risk students).

The RF classifier and the sequence model trained on the original dataset performed poorly for early prediction. This supports our hypothesis that training the model only on full traces might introduce challenges when making early predictions and highlights the effectiveness of our novel data augmentation approach for mitigating this issue.

The RF classifier with augmented data, the sequence model with augmented data, and the chapter ensemble model demonstrated similar performance for early chapters. The ROC AUC achieved a value of around 0.7 in Chapter 2, indicating moderate predictive power, and continued to improve in subsequent chapters. For a detailed report on metrics other than ROC AUC, see Appendix A.

Because early prediction is especially valuable for identifying at-risk students and informing future interventions, we also display the ROC AUC for Class 2 (at-risk students) in Figure 7 (right). The observed trend in model comparison for Class 2 closely mirrored those seen in the overall model performance.

6.2. Testing Model Generalization Across Institutions

Because the ability to generalize across institutions is a central challenge in predictive learning analytics, we next tested how well the models performed in a new institutional context. This evaluation served as a rigorous test of generalizability, simulating a real-world deployment scenario where the training and testing populations differ in institutions, instructors, student demographics, course structure, and pacing.

We trained all five model variants on data from Course 1 and evaluated their performance on an entirely new cohort of students in Course 2. Figure 8, left, shows the overall ROC AUC for this cross-institution generalization task.

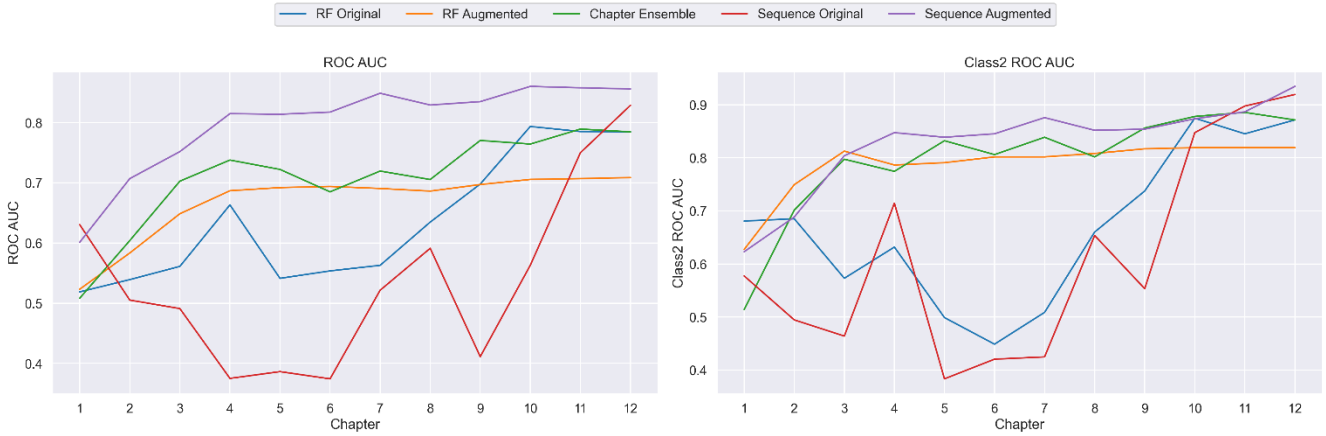

Figure 8. ROC AUC Scores in Cross-Institution Context. Note that the left graph shows the overall ROC AUC for all classes and the right graph shows the ROC AUC for Class 2 (at-risk students).

The model that emerged as noticeably superior in the generalization task was the sequence model trained on augmented data, which consistently outperformed other models starting at Chapter 2. Notably, it achieved a ROC AUC score comparable to the best-performing models on within-institution data and surpassed a ROC AUC score of 0.8 by the end of Chapter 4, indicating strong predictive performance early in the course. This suggests that capturing the evolving patterns of reading behavior and formative assessment performance across chapters is particularly valuable when predicting outcomes for a different student population. For a detailed report for metrics other than ROC AUC, see Appendix B.

When focusing specifically on at-risk students (Figure 8, right), the random forest classifier with augmented data, the sequence model with augmented data, and the chapter ensemble model all demonstrated generalization performance comparable to that observed in the overall dataset.

In sum, these findings demonstrate the promise of developing predictive models based on standardized behavioral features from a shared online textbook. Such models show strong generalizability across diverse contexts (i.e., different institutions, instructors, and course structures) and offer a potentially scalable solution for supporting student success at scale.

7. Discussion

This case study set out to evaluate the feasibility of using behavioral-only metrics derived from students' interactions with an online textbook to predict final course grades early in an introductory college-level statistics course, and to test whether such models generalize well across different institutional contexts. Our findings demonstrate not only the predictive potential and cross-institutional generalizability of the early prediction models but also highlight the importance of designing learning environments and data pipelines that support generalizable learning analytics research.

Among five predictive model variants, the sequence model trained on augmented data achieved the strongest performance, particularly for early prediction and when applied to a new student population at a different institution. Notably, by the end of Chapter 4 (approximately one-third into the course), the model trained on one institution and tested on another surpassed a ROC AUC score of 0.8. This indicates moderate to high predictive accuracy and suggests that early behavioral indicators from a shared online textbook can effectively identify students at risk of poor outcomes, even in a different learning context (i.e., different institution, instructor, and course length).

While prior studies have demonstrated promising early prediction within single institutions (e.g., Adnan et al., 2021; Chen et al., 2019; Chen & Cui, 2020; Figueroa-Cañas & Sancho-Vinuesa, 2020; Gray & Perkins, 2019; Marbouti et al., 2016), to

our knowledge, no existing research has explicitly trained a model to predict student learning outcomes for a specific course using data from one institution and evaluated it using data from a different one, particularly across varying instructional timelines, to demonstrate its generalizability. This is unsurprising, as validating model generalizability, or even developing predictive models that work across institutions, remains a persistent challenge in the learning analytics field. Even when the same course is taught in different settings, variation in instructors, student demographics, course assignments, and assessment practices can all influence model performance. Additionally, the availability and consistency of input features often differ across contexts, making it difficult to apply models developed for one class to another. For example, a model that relies heavily on historical grades may not be usable in courses that do not collect or provide access to such data. More importantly, researchers often do not have access to data beyond their own institution.

Our work directly addresses this research gap by providing an empirical demonstration of testing cross-institutional generalizability in predictive learning analytics. We leveraged CourseKata—a continuously improving, instrumented online textbook deployed across multiple institutions—to generate standardized, comparable learning data. In other words, rather than innovating on the algorithmic side, we leveraged CourseKata to create a more modelable world that would support the development of generalizable models.

Generalizability and scalability are deeply intertwined: a model that generalizes well is more likely to scale, and scalable systems provide the necessary diversity to test generalizability. The cross-institutional results from our early prediction models suggest that shared, research-aligned online learning platforms can support both goals simultaneously. A promising future direction involves scaling interventions based on these early predictions, enabling targeted support for at-risk students across varied learning environments.

Despite encouraging results, several limitations warrant consideration. First, while our models generalized across two institutions, both were public universities in California. Future work should pursue broader validation across different types of institutions (e.g., community colleges, private universities, international contexts) and diverse student populations to confirm the generalizability of findings at scale. Second, although our behavioral features provide a rich snapshot of engagement, they do not explicitly capture internal, unobservable variables such as students' motivation or beliefs. CourseKata does embed self-report surveys on these factors at multiple points in the course. Future work could explore integrating these psychological variables with behavioral traces to enhance model validity and interpretability. Third, while this study focused on final grade prediction, other outcomes (e.g., dropout, conceptual mastery, and affective states) may also benefit from early detection and intervention. Future studies could explore the predictive power and generalizability of learning data across a broader range of outcomes.

8. Conclusion

To conclude, this study proposed the development of a modelable world as a foundation for achieving generalizable and scalable predictive modeling in learning analytics and offered an empirical case study to demonstrate its effectiveness. Our findings show that behavioral data from a shared, instrumented online textbook can be used to build models that not only make accurate early predictions but also generalize across institutional contexts. We argue that the future of learning analytics should depend not only on the development of more powerful algorithms but also on the intentional design of learning environments that generate standardized, comparable, and meaningful data. By aligning research, practice, and technology within a coherent ecosystem, we can move toward predictive tools that are not just effective in isolated settings but impactful and equitable at scale.

Declaration of Conflicting Interest

Two of the authors have professional roles connected to CourseKata, a nonprofit research and curriculum development project fiscally sponsored by the National Center for Civic Innovations (NCCI). The third author serves as Chief Technology Officer at CourseKata, and the fourth author is a co-founder who contributed to the development of the curriculum used in this study. The authors have no financial or proprietary interests in any material discussed in this article.

Funding

The publication of this article received financial support from the Bill and Melinda Gates Foundation and the Valhalla Foundation.

Data Availability

The feature tables generated from the raw traces are available on GitHub at https://github.com/4lic3X/early_prediction/tree/main/data/processed. The full analysis notebooks are available on GitHub at

https://github.com/4lic3X/early_prediction/tree/main/prediction_v6. The raw traces supporting the findings of this study are not publicly available but can be obtained from CourseKata (https://www.coursekata.org/research) upon reasonable request.

References

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., ... & Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access, 9*, 7519–7539. https://doi.org/10.1109/ACCESS.2021.3049446

Alomar, K., Aysel, H. I., & Cai, X. (2023). Data augmentation in classification and segmentation: A survey and new strategies. *Journal of Imaging, 9*(2), 46. https://doi.org/10.3390/jimaging9020046

Asafuddoula, M., Verma, B., & Zhang, M. (2017). A divide-and-conquer-based ensemble classifier learning by means of many-objective optimization. *IEEE Transactions on Evolutionary Computation, 22*(5), 762–777. https://doi.org/10.1109/TEVC.2017.2782826

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*, 281–305. https://www.jmlr.org/papers/v13/bergstra12a.html

Blake, A. B., & Stigler, J. W. (2021). Track student engagement on webpages. https://uclatall.github.io/engauge

Brooks, C., Thompson, C., & Teasley, S. (2015). A time series interaction analysis method for building predictive models of learners using log data. In *LAK'15: Proceedings of the Fifth International Conference on Learning Analytics and Knowledg*e (pp. 126–135). Association for Computing Machinery. https://doi.org/10.1145/2723576.2723581

Brooks, C., & Thompson, C. (2017). Predictive Modelling in Teaching and Learning. In Lang, C., Siemens, G., Wise, A. F., and Gaevic, D., editors, *The Handbook of Learning Analytics, pages 61–68. Society for Learning Analytics Research (SoLAR)*, Alberta. https://doi.org/10.18608/hla17.005

Bulut, O., Gorgun, G., Yildirim-Erbasli, S. N., Wongvorachan, T., Daniels, L. M., Gao, Y., Lai, K. W., & Shin, J. (2023). Standing on the shoulders of giants: Online formative assessments as the foundation for predictive learning analytics models. *British Journal of Educational Technology, 54*(1), 19–39. https://doi.org/10.1111/bjet.13276

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. https://doi.org/10.1613/jair.953

Chen, W., Brinton, C. G., Cao, D., Mason-Singh, A., Lu, C., & Chiang, M. (2018). Early detection prediction of learning outcomes in online short-courses via learning behaviors. *IEEE Transactions on Learning Technologies, 12*(1), 44–58. https://doi.org/10.1109/TLT.2018.2793193

Chen, F., & Cui, Y. (2020). Utilizing student time series behaviour in learning management systems for early prediction of course performance. *Journal of Learning Analytics, 7*(2), 1–17. https://doi.org/10.18608/JLA.2020.72.1

Coburn, C. E., & Penuel, W. R. (2016). Research–practice partnerships in education: Outcomes, dynamics, and open questions. *Educational Researcher, 45*(1), 48–54. https://doi.org/10.3102/0013189X16631750

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2016). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies, 10*(1), 17–29. https://doi.org/10.1109/TLT.2016.2616312

Conole, G., Gašević, D., Long, P., & Siemens, G., (2011). Front matter (Cover, Message from the chairs, Committees, Sponsors, TOC). In *LAK'11: Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. https://doi.org/10.1145/2090116

Dawson, S., Joksimovic, S., Poquet, O., & Siemens, G. (2019, March). Increasing the impact of learning analytics. In *LAK'19: Proceedings of the First International Conference on Learning Analytics and Knowledge*. (pp. 446–455). Association for Computing Machinery. https://doi.org/10.1145/3303772.3303784

Deho, O. B., Joksimovic, S., Li, J., Zhan, C., Liu, J., & Liu, L. (2022). Should learning analytics models include sensitive attributes? Explaining the why. *IEEE Transactions on Learning Technologies, 16*(4), 560–572. https://doi.org/10.1109/TLT.2022.3226474

DiNapoli, J., & Miller, E. K. (2022). Recognizing, supporting, and improving student perseverance in mathematical problem-solving: The role of conceptual thinking scaffolds. *The Journal of Mathematical Behavior, 66*, 100965. https://doi.org/10.1016/j.jmathb.2022.100965

Fan, Y., van der Graaf, J., Lim, L. et al. Towards investigating the validity of measurement of self-regulated learning based on trace data. *Metacognition Learning 17*, 949–987 (2022). https://doi.org/10.1007/s11409-022-09291-1

Farley, I. A., & Burbules, N. C. (2022). Online education viewed through an equity lens: Promoting engagement and success for all learners. *Review of Education, 10*(3), e3367. https://doi.org/10.1002/rev3.3367

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Figueroa-Cañas, J., & Sancho-Vinuesa, T. (2020). Early prediction of dropout and final exam performance in an online statistics course. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 15*(2), 86–94. https://doi.org/10.1109/RITA.2020.2987727

Fisher, R. A. (1938). Presidential Address. *Sankhyā: The Indian Journal of Statistics (1933-1960), 4*(1), 14–17. http://www.jstor.org/stable/40383882

FIRAT, M. (2016). Determining the Effects of LMS Learning Behaviors on Academic Achievement in a Learning Analytic Perspective. *Journal of Information Technology Education Research, 15*, 75–87. https://doi.org/10.28945/3405

Fries, L., Son, J. Y., Givvin, K. B., & Stigler, J. W. (2021). Practicing connections: A framework to guide instructional design for developing understanding in complex domains. *Educational Psychology Review, 33*(2), 739–762. https://doi.org/10.1007/s10648-020-09561-x

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education, 28*, 68–84. https://doi.org/10.1016/j.iheduc.2015.10.002

Ge, L., Gao, J., Ngo, H., Li, K., & Zhang, A. (2014). On handling negative transfer and imbalanced distributions in multiple source transfer learning. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 7*(4), 254–271. https://doi.org/10.1002/sam.11217

Gitinabard, N., Xu, Y., Heckman, S., Barnes, T., & Lynch, C. F. (2019). How widely can prediction models be generalized? Performance prediction in blended courses. *IEEE Transactions on Learning Technologies, 12*(2), 184–197. https://doi.org/10.1109/TLT.2019.2911832

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics, 21*(2), 215-223. https://doi.org/10.1080/00401706.1979.10489751

Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers & Education, 131*, 22–32. https://doi.org/10.1016/j.compedu.2018.12.006

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hernández-García, Á., Cuenca-Enrique, C., Del-Río-Carazo, L., & Iglesias-Pradas, S. (2024). Exploring the relationship between LMS interactions and academic performance: A Learning Cycle approach. *Computers in Human Behavior, 155*(2024). 108183. https://doi.org/10.1016/j.chb.2024.108183

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hossin, M., & Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process, 5*(2), 1. https://doi.org/10.5121/ijdkp.2015.5201

Hu, Y. H. (2022). Effects of the COVID-19 pandemic on the online learning behaviors of university students in Taiwan. *Education and Information Technologies, 27*(1), 469–491. https://doi.org/10.1007/s10639-021-10677-y

Huang, A. Y. Q., Lu, O. H. T., Huang, J. C. H., Yin, C. J., & Yang, S. J. H. (2019). Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. *Interactive Learning Environments, 28*(2), 206–230. https://doi.org/10.1080/10494820.2019.1636086

Ifenthaler, D. (2017). Are higher education institutions prepared for learning analytics? *TechTrends, 61*(4), 366–371. https://doi.org/10.1007/s11528-016-0154-0

Junco, R., & Clem, C. (2015). Predicting course outcomes with digital textbook usage data. *The Internet and Higher Education, 27*, 54–63. https://doi.org/10.1016/j.iheduc.2015.06.001

Kim, D., Park, Y., Yoon, M., & Jo, I. H. (2016). Toward evidence-based learning analytics: Using proxy variables to improve asynchronous online discussion environments. *The Internet and Higher Education, 30*, 30–43. https://doi.org/10.1016/j.iheduc.2016.03.002

Kim, D., Yoon, M., Jo, I.-H., & Branch, R. M. (2018). Learning analytics to support self-regulated learning in asynchronous online courses: A case study at a women's university in South Korea. *Computers & Education, 127*, 233–251. https://doi.org/10.1016/j.compedu.2018.08.023

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2,* 1137–1143. Presented at the Montreal, Quebec, Canada. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R. S., & Hatala, M. (2015). Penetrating the black box of time-on-task estimation. In *LAK '15: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*. (pp.184–193.) Association for Computing Machinery. https://doi.org/10.1145/2723576.2723623

Langley, G. J., Moen, R. D., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide: a practical approach to enhancing organizational performance*. John Wiley & Sons.

Lewis, C. (2015). What is improvement science? Do we need it in education? *Educational Researcher, 44*(1), 54–61. https://doi.org/10.3102/0013189X15570388

Lidwell, W., Holden, K., & Butler, J. (2010). *Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design.* Rockport Pub.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue, 16*(3), 31–57.

Liu, Z., He, J., Xue, Y., Huang, Z., Li, M., & Du, Z. (2015). Modeling the learning behaviors of massive open online courses. *2015 IEEE International Conference on Big Data*, 2883–2885. https://doi.org/10.1109/BigData.2015.7364110

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings, 3*(1), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., & Munoz-Gama, J. (2018). Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses. *Computers in Human Behavior, 80*, 179–196. https://doi.org/10.1016/j.chb.2017.11.011

Marbouti, Farshid, Heidi A. Diefes-Dux, and Krishna Madhavan. "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

Mathrani, A., Susnjak, T., Ramaswami, G., & Barczak, A. (2021). Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. *Computers and Education Open, 2*, 100060. https://doi.org/10.1016/j.caeo.2021.100060

Mega, C., Ronconi, L., & De Beni, R. (2014). What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement. *Journal of Educational Psychology, 106*(1), 121–131. https://doi.org/10.1037/a0033546

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*, 5–11. https://doi.org/10.3102/0013189X018002005

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: overview study and experimental results. In I*nternational Conference on Information and Communication Systems (ICICS)* (pp. 243–248). IEEE.

Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., & Kloos, C. D. (2018). Prediction in MOOCs: A review and future research directions. *IEEE transactions on Learning Technologies, 12*(3), 384–401. https://doi.org/10.1109/TLT.2018.2856808

Munir, H., Moayyed, M., & Petersen, K. (2014). Considering rigor and relevance when evaluating test driven development: A systematic review. I*nformation and Software Technology, 56*(4), 375–394. https://doi.org/10.1016/j.infsof.2014.01.002

Mwalumbwe, I., & Mtebe, J. S. (2017). Using learning analytics to predict students' performance in Moodle learning management system: A case of Mbeya University of Science and Technology. *The Electronic Journal of Information Systems in Developing Countries, 79*(1), 1–13. https://doi.org/10.1002/j.1681-4835.2017.tb00577.x

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://doi.org/10.48550/arXiv.1201.0490

Pelánek, R. (2020, March). Learning analytics challenges: trade-offs, methodology, scalability. In *LAK'20: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. (pp. 554–558). Association for Computing Machinery. https://doi.org/10.1145/3375462.33754

Provost, F., & Domingos, P. (2000). Well-trained PETs: Improving probability estimation trees. *Technical Report IS-00-04*, Stern School of Business, New York University.

Samuelsen, J., Chen, W. & Wasson, B. (2019). Integrating multiple data sources for learning analytics—review of literature. *Research and Practice in Technology Enhanced Learning, 13*(11). https://doi.org/10.1186/s41039-019-0105-4

Sghir, N., Adadi, A., & Lahmer, M. (2023). Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022). *Education and Information Technologies, 28*(7), 8299–8333. https://doi.org/10.1007/s10639-022-11536-0

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing, 97*, 105524. https://doi.org/10.1016/j.asoc.2019.105524

Sharma, R., Shen, H., & Goodwin, R. (2016). Voluntary participation in discussion forums as an engagement indicator: an empirical study of teaching first-year programming. In *OzCHI '16: Proceedings of the 28th Australian Conference on Computer-Human Interaction*. (pp. 489–493). Association for Computing Machinery. https://doi.org/10.1145/3010915.3010967

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior, 98*, 277–284. https://doi.org/10.1016/j.chb.2019.04.019

Son, J. Y., & Stigler, J. W. (2017-2024). *Data Science and Statistics: A Modeling Approach.* https://coursekata.org/preview/default/program

Stigler, J. W., Son, J. Y., Givvin, K. B., Blake, A. B., Fries, L., Shaw, S. T., & Tucker, M. C. (2020). The Better Book Approach for Education Research and Development. *Teachers College Record, 122*(9), 1–32. https://doi.org/10.1177/016146812012200913

Storkey, A. (2009). When training and test sets are different: characterizing learning transfer. In Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D., *Dataset Shift in Machine Learning,* (pp. 2–28), MIT Press. https://doi.org/10.7551/mitpress/9780262170055.003.0001

Suleman, S., Antu, S. W., & Malanua, S. (2022). The Influence of Assignment Methods and Learning Behavior on Student Learning Outcomes. *Journal La Edusci, 3*(2), 28–36. https://doi.org/10.37899/journallaedusci.v3i2.607

Sutter, C. C., Totonchi, D. A., DeCoster, J., Barron, K. E., & Hulleman, C. S. (2024). How does expectancy-value-cost motivation vary during a semester? An intensive longitudinal study to explore individual and situational sources of variation in statistics motivation. *Learning and Individual Differences, 113*, 102484. https://doi.org/10.1016/j.lindif.2024.102484

Tate, T., & Warschauer, M. (2022). Equity in online learning. *Educational Psychologist, 57*(3), 192–206. https://doi.org/10.1080/00461520.2022.2062597

Trautwein, U., Niggli, A., Schnyder, I., & Lüdtke, O. (2009). Between-teacher differences in homework assignments and the development of students' homework effort, homework emotions, and achievement. *Journal of Educational Psychology, 101*(1), 176–189. https://doi.org/10.1037/0022-0663.101.1.176

Tricot, A., & Sweller, J. (2014). Domain-specific knowledge and why teaching generic skills does not work. *Educational Psychology Review, 26*(2), 265–283. https://doi.org/10.1007/s10648-013-9243-1

Tsai, Y. S., Rates, D., Moreno-Marcos, P. M., Munoz-Merino, P. J., Jivet, I., Scheffel, M., ... & Gašević, D. (2020). Learning analytics in European higher education—Trends and barriers. *Computers & Education, 155*, 103933. https://doi.org/10.1016/j.compedu.2020.103933

Tucker, M. C., Shaw, S. T., Son, J. Y., & Stigler, J. W. (2023). Teaching statistics and data analysis with R. *Journal of Statistics and Data Science Education, 31*(1), 18–32. https://doi.org/10.1080/26939169.2022.2089410

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE, 14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

van Halem, N., van Klaveren, C., Drachsler, H., Schmitz, M., & Cornelisz, I. (2020). Tracking patterns in self-regulated learning using students' self-reports and online trace data. *Frontline Learning Research, 8*(3), 140–163. https://doi.org/10.14786/flr.v8i3.497

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace.

Wilson, A., Watson, C., Thompson, T. L., Drew, V., & Doyle, S. (2017). Learning analytics: Challenges and limitations. *Teaching in Higher Education, 22*(8), 991–1007. https://doi.org/10.1080/13562517.2017.1332026

Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist, 45*, 267–276. https://doi.org/10.1080/00461520.2010.517150.

Winne, P. H. (2020). Construct and consequential validity for learning analytics based on trace data. *Computers in Human Behavior, 112*, 106457. https://doi.org/10.1016/j.chb.2020.106457

Xu, A., Blake, A. B., Zhang, I. Y., Zhao, Y., & Epner, R. (2023). Early Identification of Underperforming Students via Reading Patterns. In *American Educational Research Association Online Paper Repository*. https://doi.org/10.3102/2009101

Yang, T. Y., Brinton, C. G., Joe-Wong, C., & Chiang, M. (2017). Behavior-based grade prediction for MOOCs via time series neural networks. *IEEE Journal of Selected Topics in Signal Processing, 11*(5), 716–728. https://doi.org/10.1109/JSTSP.2017.2700227

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series, 1168*(2). https://doi.org/10.1088/1742-6596/1168/2/022022

Zimmerman, T. D. (2012). Exploring learner to content interaction as a success factor in online courses. *International Review of Research in Open and Distance Learning, 13*(4). https://www.proquest.com/scholarly-journals/exploring-learner-content-interaction-as-success/docview/1634473616/se-2

Appendix A

Table 1. Metrics for RF Augmented Versus RF Original Model Within Course 1

RF Augmented vs. RF Original

Chapter	1	2	3	4	5	6	7	8	9	10	11	12
ROC_AUC	0.59 (0.52)	0.69 (0.58)	0.7 (0.6)	0.72 (0.59)	0.75 (0.57)	0.76 (0.56)	0.76 (0.57)	0.77 (0.68)	0.79 (0.79)	0.81 (0.84)	0.81 (0.85)	0.81 (0.86)
Accuracy	0.57 (0.29)	0.65 (0.34)	0.66 (0.39)	0.68 (0.43)	0.69 (0.46)	0.7 (0.48)	0.7 (0.49)	0.71 (0.56)	0.71 (0.68)	0.71 (0.7)	0.71 (0.72)	0.71 (0.73)
Precision	0.55 (0.16)	0.6 (0.34)	0.6 (0.47)	0.6 (0.46)	0.62 (0.47)	0.63 (0.48)	0.65 (0.51)	0.65 (0.6)	0.65 (0.7)	0.65 (0.7)	0.65 (0.73)	0.65 (0.72)
Recall	0.57 (0.29)	0.65 (0.34)	0.66 (0.39)	0.68 (0.43)	0.69 (0.46)	0.7 (0.48)	0.7 (0.49)	0.71 (0.56)	0.71 (0.68)	0.71 (0.7)	0.71 (0.72)	0.71 (0.73)
F1	0.55 (0.17)	0.62 (0.26)	0.62 (0.34)	0.63 (0.39)	0.64 (0.42)	0.66 (0.44)	0.66 (0.45)	0.67 (0.53)	0.67 (0.67)	0.67 (0.69)	0.67 (0.72)	0.67 (0.72)
Class2_ROC_AUC	0.58 (0.51)	0.72 (0.54)	0.72 (0.54)	0.73 (0.54)	0.76 (0.5)	0.78 (0.47)	0.78 (0.46)	0.79 (0.68)	0.81 (0.81)	0.82 (0.85)	0.82 (0.86)	0.82 (0.86)
Class2_Precision	0.37 (0.19)	0.52 (0.21)	0.54 (0.22)	0.55 (0.24)	0.6 (0.22)	0.61 (0.24)	0.62 (0.22)	0.64 (0.37)	0.63 (0.57)	0.62 (0.54)	0.62 (0.72)	0.62 (0.73)
Class2_Recall	0.32 (0.77)	0.47 (0.72)	0.41 (0.63)	0.43 (0.59)	0.46 (0.51)	0.47 (0.4)	0.48 (0.34)	0.47 (0.51)	0.48 (0.64)	0.49 (0.69)	0.49 (0.59)	0.49 (0.61)
Class2_F1	0.33 (0.3)	0.49 (0.32)	0.46 (0.32)	0.47 (0.32)	0.51 (0.29)	0.52 (0.26)	0.52 (0.23)	0.53 (0.39)	0.53 (0.57)	0.53 (0.6)	0.53 (0.64)	0.53 (0.65)

Table 2. Metrics for Sequence Augmented Versus Sequence Original Model Within Course 1

Sequence Augmented vs. Sequence Original

Chapter	1	2	3	4	5	6	7	8	9	10	11	12
ROC_AUC	0.61 (0.55)	0.7 (0.56)	0.72 (0.57)	0.74 (0.59)	0.76 (0.62)	0.77 (0.66)	0.78 (0.69)	0.8 (0.73)	0.83 (0.76)	0.86 (0.83)	0.87 (0.85)	0.86 (0.88)
Accuracy	0.45 (0.33)	0.54 (0.35)	0.6 (0.35)	0.63 (0.34)	0.66 (0.35)	0.67 (0.4)	0.68 (0.44)	0.7 (0.53)	0.72 (0.61)	0.73 (0.67)	0.73 (0.68)	0.72 (0.71)
Precision	0.59 (0.24)	0.64 (0.25)	0.66 (0.28)	0.66 (0.29)	0.69 (0.35)	0.7 (0.44)	0.71 (0.55)	0.73 (0.64)	0.76 (0.69)	0.76 (0.74)	0.79 (0.77)	0.78 (0.8)
Recall	0.45 (0.33)	0.54 (0.35)	0.6 (0.35)	0.63 (0.34)	0.66 (0.35)	0.67 (0.4)	0.68 (0.44)	0.7 (0.53)	0.72 (0.61)	0.73 (0.67)	0.73 (0.68)	0.72 (0.71)
F1	0.44 (0.24)	0.55 (0.26)	0.61 (0.27)	0.64 (0.26)	0.66 (0.28)	0.67 (0.35)	0.69 (0.42)	0.7 (0.54)	0.73 (0.62)	0.74 (0.69)	0.75 (0.7)	0.74 (0.73)
Class2_ROC_AUC	0.61 (0.55)	0.71 (0.57)	0.75 (0.59)	0.75 (0.6)	0.77 (0.62)	0.8 (0.67)	0.8 (0.7)	0.83 (0.76)	0.85 (0.77)	0.86 (0.84)	0.87 (0.86)	0.87 (0.88)
Class2_Precision	0.54 (0.07)	0.59 (0.12)	0.61 (0.11)	0.58 (0.12)	0.61 (0.22)	0.65 (0.47)	0.65 (0.47)	0.72 (0.56)	0.72 (0.68)	0.74 (0.72)	0.78 (0.76)	0.74 (0.75)
Class2_Recall	0.23 (0.14)	0.41 (0.12)	0.46 (0.13)	0.48 (0.15)	0.5 (0.16)	0.53 (0.23)	0.54 (0.29)	0.57 (0.41)	0.57 (0.48)	0.61 (0.55)	0.63 (0.57)	0.63 (0.64)
Class2_F1	0.28 (0.09)	0.45 (0.09)	0.5 (0.09)	0.51 (0.11)	0.53 (0.14)	0.56 (0.26)	0.58 (0.32)	0.62 (0.45)	0.62 (0.54)	0.66 (0.61)	0.69 (0.63)	0.67 (0.67)

Table 3. Metrics for Chapter Ensemble Model Within Course 1

Chapter Ensemble

Chapter	1	2	3	4	5	6	7	8	9	10	11	12
ROC_AUC	0.6	0.7	0.7	0.7	0.73	0.74	0.75	0.78	0.84	0.85	0.86	0.86
Accuracy	0.58	0.64	0.64	0.64	0.68	0.7	0.68	0.7	0.72	0.73	0.73	0.73
Precision	0.56	0.62	0.61	0.62	0.66	0.69	0.66	0.69	0.72	0.72	0.73	0.72
Recall	0.58	0.64	0.64	0.64	0.68	0.7	0.68	0.7	0.72	0.73	0.73	0.73
F1	0.56	0.62	0.61	0.62	0.66	0.68	0.66	0.69	0.71	0.72	0.72	0.72
Class2_ROC_AUC	0.58	0.7	0.71	0.7	0.73	0.76	0.76	0.83	0.87	0.86	0.86	0.86
Class2_Precision	0.41	0.48	0.51	0.56	0.61	0.68	0.64	0.77	0.73	0.7	0.7	0.73
Class2_Recall	0.33	0.48	0.44	0.47	0.45	0.48	0.47	0.55	0.58	0.59	0.59	0.61
Class2_F1	0.35	0.47	0.46	0.5	0.5	0.55	0.52	0.63	0.63	0.63	0.63	0.65

Appendix B

Table 1. Metrics for RF Augmented Versus RF Original Model from Course 1 to Course 2

RF Augmented vs. RF Original

Chapter	1	2	3	4	5	6	7	8	9	10	11	12
ROC_AUC	0.52 (0.52)	0.58 (0.54)	0.65 (0.56)	0.69 (0.66)	0.69 (0.54)	0.69 (0.55)	0.69 (0.56)	0.69 (0.63)	0.7 (0.7)	0.71 (0.79)	0.71 (0.79)	0.71 (0.78)
Accuracy	0.45 (0.17)	0.58 (0.17)	0.68 (0.55)	0.72 (0.7)	0.73 (0.4)	0.72 (0.65)	0.72 (0.58)	0.72 (0.65)	0.7 (0.68)	0.7 (0.77)	0.7 (0.73)	0.7 (0.73)
Precision	0.55 (0.05)	0.61 (0.39)	0.57 (0.61)	0.61 (0.65)	0.63 (0.47)	0.61 (0.57)	0.61 (0.57)	0.61 (0.59)	0.61 (0.58)	0.61 (0.67)	0.61 (0.65)	0.61 (0.65)
Recall	0.45 (0.17)	0.58 (0.17)	0.68 (0.55)	0.72 (0.7)	0.73 (0.4)	0.72 (0.65)	0.72 (0.58)	0.72 (0.65)	0.7 (0.68)	0.7 (0.77)	0.7 (0.73)	0.7 (0.73)
F1	0.49 (0.08)	0.57 (0.13)	0.62 (0.58)	0.66 (0.66)	0.68 (0.43)	0.66 (0.59)	0.66 (0.57)	0.66 (0.61)	0.65 (0.63)	0.65 (0.7)	0.65 (0.67)	0.65 (0.67)
Class2_ROC_AUC	0.63 (0.68)	0.75 (0.69)	0.81 (0.57)	0.79 (0.63)	0.79 (0.5)	0.8 (0.45)	0.8 (0.51)	0.81 (0.66)	0.82 (0.74)	0.82 (0.87)	0.82 (0.85)	0.82 (0.87)
Class2_Precision	0.4 (0.2)	0.5 (0.21)	0.5 (0.31)	0.5 (0.4)	0.5 (0.07)	0.5 (0.2)	0.5 (0.17)	0.5 (0.29)	0.5 (0.45)	0.5 (1.0)	0.5 (1.0)	0.5 (1.0)
Class2_Recall	0.22 (0.78)	0.11 (0.44)	0.33 (0.44)	0.67 (0.22)	0.67 (0.11)	0.67 (0.11)	0.67 (0.22)	0.67 (0.22)	0.67 (0.56)	0.67 (0.67)	0.67 (0.56)	0.67 (0.56)
Class2_F1	0.29 (0.32)	0.18 (0.29)	0.4 (0.36)	0.57 (0.29)	0.57 (0.08)	0.57 (0.14)	0.57 (0.19)	0.57 (0.25)	0.57 (0.5)	0.57 (0.8)	0.57 (0.71)	0.57 (0.71)

Table 2. Metrics for Sequence Augmented Versus Sequence Original Model from Course 1 to Course 2

Sequence Augmented vs. Sequence Original

Chapter	1	2	3	4	5	6	7	8	9	10	11	12
ROC_AUC	0.6 (0.63)	0.71 (0.51)	0.75 (0.49)	0.81 (0.38)	0.81 (0.39)	0.82 (0.37)	0.85 (0.52)	0.83 (0.59)	0.83 (0.41)	0.86 (0.56)	0.86 (0.75)	0.86 (0.83)
Accuracy	0.28 (0.2)	0.65 (0.2)	0.65 (0.18)	0.67 (0.2)	0.7 (0.18)	0.72 (0.15)	0.72 (0.17)	0.72 (0.22)	0.75 (0.25)	0.77 (0.32)	0.77 (0.77)	0.78 (0.75)
Precision	0.63 (0.71)	0.67 (0.74)	0.66 (0.06)	0.7 (0.31)	0.68 (0.06)	0.71 (0.18)	0.71 (0.06)	0.69 (0.75)	0.74 (0.52)	0.74 (0.64)	0.75 (0.76)	0.77 (0.72)
Recall	0.28 (0.2)	0.65 (0.2)	0.65 (0.18)	0.67 (0.2)	0.7 (0.18)	0.72 (0.15)	0.72 (0.17)	0.72 (0.22)	0.75 (0.25)	0.77 (0.32)	0.77 (0.77)	0.78 (0.75)
F1	0.3 (0.14)	0.66 (0.11)	0.65 (0.08)	0.68 (0.15)	0.67 (0.09)	0.7 (0.1)	0.7 (0.09)	0.7 (0.15)	0.72 (0.21)	0.74 (0.28)	0.73 (0.75)	0.75 (0.71)
Class2_ROC_AUC	0.62 (0.58)	0.69 (0.49)	0.8 (0.46)	0.85 (0.71)	0.84 (0.38)	0.85 (0.42)	0.88 (0.42)	0.85 (0.65)	0.85 (0.55)	0.87 (0.85)	0.89 (0.9)	0.93 (0.92)
Class2_Precision	0.4 (0.0)	0.5 (0.2)	0.46 (0.17)	0.4 (0.33)	0.43 (0.17)	0.46 (0.14)	0.46 (0.14)	0.55 (0.18)	0.6 (0.2)	0.75 (1.0)	0.67 (0.83)	0.78 (1.0)
Class2_Recall	0.22 (0.0)	0.44 (0.11)	0.67 (0.22)	0.67 (0.11)	0.67 (0.33)	0.67 (0.33)	0.67 (0.67)	0.67 (0.89)	0.67 (0.67)	0.67 (0.56)	0.67 (0.56)	0.78 (0.56)
Class2_F1	0.29 (0.0)	0.47 (0.14)	0.55 (0.19)	0.5 (0.17)	0.52 (0.22)	0.55 (0.19)	0.55 (0.23)	0.6 (0.3)	0.63 (0.31)	0.71 (0.71)	0.67 (0.67)	0.78 (0.71)

Table 3. Metrics for Chapter Ensemble Model from Course 1 to Course 2

Chapter Ensemble

Chapter	1	2	3	4	5	6	7	8	9	10	11	12
ROC_AUC	0.51	0.6	0.7	0.74	0.72	0.68	0.72	0.71	0.77	0.76	0.79	0.78
Accuracy	0.53	0.58	0.68	0.65	0.68	0.7	0.72	0.73	0.75	0.73	0.73	0.73
Precision	0.55	0.57	0.58	0.61	0.58	0.61	0.68	0.63	0.72	0.65	0.65	0.65
Recall	0.53	0.58	0.68	0.65	0.68	0.7	0.72	0.73	0.75	0.73	0.73	0.73
F1	0.54	0.57	0.63	0.62	0.63	0.65	0.67	0.68	0.71	0.67	0.67	0.67
Class2_ROC_AUC	0.51	0.7	0.8	0.77	0.83	0.81	0.84	0.8	0.86	0.88	0.89	0.87
Class2_Precision	0.25	0.6	0.45	0.38	0.38	0.55	0.62	0.67	1.0	1.0	1.0	1.0
Class2_Recall	0.22	0.33	0.56	0.67	0.56	0.67	0.56	0.67	0.56	0.56	0.56	0.56
Class2_F1	0.24	0.43	0.5	0.48	0.45	0.6	0.59	0.67	0.71	0.71	0.71	0.71