Summer 2022

Predicting Students’ Performance Using Growth Models

Sam Habach

SHAWNEE STATE UNIVERSITY

Predicting Students’ Performance Using Growth Models

A Thesis

By

Sam Habach

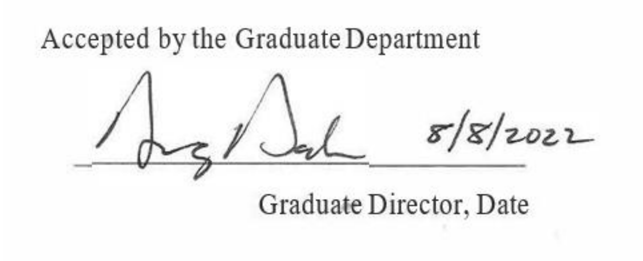
Department of Mathematical Sciences

Submitted in partial fulfillment of the requirements

for the degree of

Master of Science, Mathematics

August 8, 2022

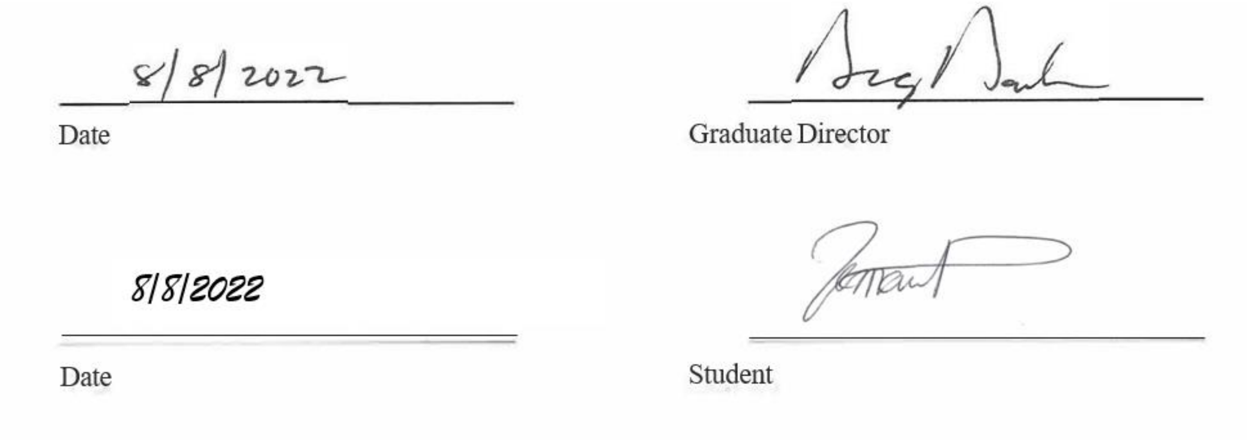


Accepted by the Graduate Department

8/8/2022

Graduate Director, Date

The thesis entitled '**Predicting Students' Performance Using Growth Models**' presented by **SAM HABACH**, a candidate for the degree of **Master of Science in Mathematics,** has been approved and is worthy of acceptance.



8/8/2022

Date

Graduate Director

8/8/2022

Date

Student

ABSTRACT

This study examines the effectiveness of latent growth models for clustering middle school math students to predict the high school academic outcome, as measured by the number of AP STEM courses passed in high school. Middle school math performance is based on the 6th, 7th, and 8th grade Rhode Island Comprehensive Assessment System (RICAS) scores. The study also examines whether demographic factors can predict student performance clusters. These demographic factors are High-risk status (HAR), eligibility for a free lunch program (FRP), and/or an individualized learning plan (IEP). The study concluded that the students' growth can be grouped into four clusters. In addition, it demonstrated that cluster membership was associated with demographic predictors consistent with expected historical trends. Furthermore, the study found that the high school performance of students can be predicted by their membership in a growth cluster.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to Dr. Shane Tutwiler from the University of Rhode Island and Dr. Douglas Darbro from Shawnee State University for their assistance at every stage of the research project.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I: INTRODUCTION

One of the challenges that educational institutions face is the lack of an analytical foundation that can be used to utilize large data to make effective and evidence-based decisions that improve students' education quality and assist educators in providing the necessary intervention to ensure that all students have equal opportunity.

Schools and districts collect a large amount of information on their students. Despite the abundance of data, many of these organizations lack the thoughtful evaluation to give accurate information and make evidence-based decisions.

The state of Rhode Island is one of several that possesses a large amount of data that hasn't been thoroughly studied. The goal of this study is to lay a foundation for the Rhode Island Department of Education to use in order to improve education quality for Rhode Island's students.

Background of the Problem

Many studies have been carried out in order to investigate models that can be used to predict students' success based on their previous performance. A study by Abu-Naser et al. found that the ANN model could accurately predict the performance of more than 80 percent of prospective students in a given semester (Abu-Naser, Samy S.; Zaqout, Ihab S.; Abu Ghosh, Mahmoud; Atallah, Rasha R.; Eman Alajrami; 2015).

Numerous studies have also conducted comparative research to determine how effectively different modeling strategies predict students' performance based on quantitative parameters. For example, one study by Sokkhery et al. conducted a comparison study between statistical analytic methodologies, machine learning (ML) algorithms, and deep learning architecture. The study discovered that the Random Forest (RF) outperformed other models in predicting student performance (Phauk Sokkhey; Takeo Okazaki; 2019).

When it comes to growth models, however, there have been few studies that have looked at how successful this tool is in predicting student achievement. Perhaps the most significant stumbling block is a scarcity of longitudinal data, which provides detailed information on the students' academic performance over time. One of these few studies is a study by Mirayama et al., where growth models were used to predict students' long-term growth in mathematics.

A large number of studies, on the other hand, have been conducted with the goal of predicting students' performance based on external factors such as socioeconomic and demographic parameters. According to Ding et al. research, over time, both male and female students showed the same development trend in mathematical performance, but female students' mathematics grade-point average was significantly higher than male students' mathematics grade-point average over the same period (Cody S. Ding; Kim Song; and Lloyd I Richardson; 2010).

Statement of the Problem

Unlike the previous study by Abu-Naser et al., which used ANN models and general clustering models, this study will use latent growth models to cluster students according to their performance in middle school mathematics and investigate the extent to which these clusters can be used to predict students' outcomes in high school. While the study by Mirayama et al. used growth models, it did not use students' previous performance as a possible predicting factor of the students' academic outcome. This information is vital for educational institutions to assess the effectiveness of the curriculum. Finally, although the previous study by Ding et al.

investigated the relationship between gender and performance in mathematics, it did not include other socioeconomic and demographic factors. This study, on the other hand, will incorporate three demographic factors that are known to pose obstacles to students and investigate the extent to which these three demographic factors can predict students' performance clusters, and hence their academic outcome in high school.

In conclusion, what distinguishes this study from others is that it attempts to use latent growth models to cluster students' performance based on their performance in middle school mathematics, investigates the extent to which demographic factors (the three demographic factors) can predict these students' performance clusters, and predicts the students' academic outcome in high school based on these performance clusters.

Purpose of the Study

This study aims to investigate the utility of latent growth models in clustering students based on their performance in middle school mathematics to eventually predict the students' academic outcome in high school. Students' performance in middle school mathematics is obtained from their scores in their 6th, 7th, and 8th grades of the Rhode Island Comprehensive Assessment System (RICAS). Students' outcomes refer to students' performance in high school. In particular, it refers to the number of AP STEM courses taken in high school, the number of AP STEM courses passed in high school, and the total number of STEM courses enrolled in high school. The study will also investigate the extent to which the three demographic factors can predict students' performance clusters. These three demographic factors are students' high-risk status (HR), eligibility for a free lunch program (FLP), and/or need for an individualized educational plan (IEP).

The study is both hypothesis testing and predictive. The latent growth models will be used to analyze the data in order to determine whether the hypotheses should be accepted or rejected. On the other hand, the analysis will result in the development of a model that can predict the students' academic outcome based on their performance clusters from their performance in middle school mathematics as well as the three demographic factors (HR, FLP, and IEP).

Significance of the Study

The significance of the study lies in the fact that it will be the first attempt to cluster students in the state of Rhode Island based on their growth pattern, which will be determined using latent growth models derived from students' middle school mathematics performance. The study will then examine the extent to which these clusters can be used as indicators of high school academic performance. The study will also examine the extent to which three demographic factors (the three demographic factors) can predict these clusters and, by extension, high school academic performance. This will enable the Rhode Island Department of Education to make informed decisions regarding the allocation of resources to curriculum development, special education programs, integration, equality, and diversity initiatives, based on reliable information. It will also provide information to other educational institutions and educators to aid in the detection of warning signs and the implementation of appropriate intervention strategies to ensure that all students have equal access to opportunities.

Primary Research Questions

Question 1: Can students be clustered based on their performance growth in middle school mathematics?

Question 2: Can the three demographic factors be used as indicators to predict students' growth performance clusters (students' high-risk status (HR), eligibility for a free lunch program (FLP), and/or need for an individualized educational plan? (IEP)?

Question 3: Can students' growth performance clusters predict the students' academic outcome in high school?

Hypotheses

Hypothesis 1: Students can be clustered using latent growth models based on their performance growth in middle school mathematics.

Hypothesis 2: The three demographic factors can be used as indicators to predict students' growth performance clusters.

Hypothesis 3: Students' growth performance clusters predict the students' academic outcome in high school.

Research Design

The research will use R, which is a free, open-source software that uses R language to perform statistical analyses. In R, latent growth models in the package *lavaan* will be used. The research will start by clustering students based on their performance. The resulting clusters will be confirmed using other clustering techniques such as SEM.

Next, the study will investigate if these performance clusters are affected when taking the district and school into account.

Next, the study will investigate if students' performance in middle school mathematics can predict students' performance in high school mathematics.

Next, the study will investigate the extent to which these performance clusters can be used to predict the students' academic outcome in high school, measured by the number of AP STEM courses passed in high school.

Next, the study will implement the three demographic factors into the latent growth models to investigate their influence on predicting students' performance in middle school mathematics.

Next, the study will investigate if the three demographic factors can be used as indicators to predict students' performance clusters, and hence their academic outcome in high school.

Once these findings are established, the research will make a conclusion on the effectiveness of using latent growth models in clustering students based on their performance which is obtained from their performance in middle school mathematics. It will also make a conclusion on the extent to which the three demographic factors can be used to predict students' performance clusters, and hence their academic outcome in high school.

Theoretical Framework

The study will attempt to cluster students based on their performance using latent growth modeling, which is a statistical approach for estimating growth trajectories inside the structural equation modeling (SEM) framework. It is a strategy for estimating growth over time using longitudinal analysis through repeated dependent variable measurements as a function of time and other factors. These longitudinal data have the common characteristic of observing the same variable repeatedly throughout time. In latent growth modeling, each variable's relative position at every point in time is represented as a function of an underlying growth process, with the optimized values for that growth process fitted to each variable.

To incorporate the three demographic factors into the model, the study will also use multinomial regression modeling, which is a classification approach that extends the scope of logistic regression to include outcomes with more than two potential discrete solutions. A categorical dependent variable with a categorically distributed probability distribution is most often used to estimate the probabilities of the many possible outcomes given a collection of independent variables.

In addition, the study will use poison regression modeling to predict the students' academic outcome in high school based on their performance cluster. Poisson regression is a kind of generalized linear model that is often used to describe count data. This approach is based on the presumption that the dependent variable follows a Poisson distribution and that the logarithm of its expected value may be described using a linear combination of parameters. This method is especially useful when the dependent variable is a count. In this study, the dependent variable, which represents students' outcomes in high school, is the number (count) of advanced math or science classes taken and passed.

Assumptions, Limitations, and Scope

The study's scope is defined by the number of AP STEM courses taken and passed. The study, therefore, makes no mention of additional factors that could be considered academic performance in high school, such as dual enrollment, honors courses, extracurricular activities, and college admission.

In addition, the study’s indicating factor is limited to students’ performance in middle school mathematics; it makes no mention of other indicating factors that could easily be considered performance factors, such as performance in non-mathematics courses or classroom interaction.

To maximize the study’s efficiency, only students with constant demographic factors throughout their middle school years were included. This resulted in a negligible decrease in student enrollment and thus had no effect on the study’s findings.

Finally, other factors such as the school district or high school may have a significant impact on what this study defines as the students’ academic outcome in high school. To address this issue, the study fitted a multinomial regression to account for such potentially confounding factors.

Definition of Terms

Performance clusters: Clusters that the study attempts to generate using latent growth models from students’ performance in middle school mathematics.

Students’ performance in middle school mathematics: Data obtained from students’ scores in their 6th, 7th, and 8th grades of the Rhode Island’s Comprehensive Assessment System (RICAS),

The students’ academic outcome in high school: The number of AP STEM courses passed in high school.

The three demographic factors: Warning signs that may hinder students’ success; whether the student is in a high-risk group (HR); whether the student is eligible for a free lunch program (FLP); and whether the student is eligible for an individualized education plan (IEP),

Summary

Chapter I introduced the problem to be investigated by providing a context for the issue from previous research papers and formulating a problem statement. Additionally, it stated the study's purpose and demonstrated its significance. Chapter I also included a list of the study's primary research questions and hypotheses. Finally, it discussed the research design and provided an overview of the study's theoretical framework, as well as the study's assumptions, limitations, and scope.

Chapter II will examine previous research that attempted to predict the future outcomes of students based on their past performances. This chapter will provide an overview of the general research methods, experiments, and procedures used in these studies, as well as data science methodology, and will highlight the uniqueness of this study and why it is significant and relevant to education research.

In Chapter III, the process and methodology of this research will be discussed in depth. There will be a summary of the research questions, how data was collected, what statistical models were used and how they were used, what programming languages or software were used to facilitate the research, and how the data was analyzed and examined.

In Chapter IV, the study's findings will be discussed, with a focus on those that are statistically significant. Next, Chapter V will evaluate the findings, highlight the study's contributions to education research, identify any limitations or gaps in the findings, and suggest future research topics that would fill the gaps and enhance the current research's findings and outcomes.

CHAPTER 2: LITERATURE REVIEW

This chapter will review the prior research and studies conducted on this paper's subject. It will begin by reviewing prior research that has used models to predict students' success based on prior performance. Following that, the review will discuss previous studies that conducted comparative analyses of the effectiveness and advantages and disadvantages of various models for predicting student performance. Thirdly, the review will shed light on previous studies that specifically used growth models to predict student performance. Fourth, the chapter will highlight previous research studies on the effects of external factors on student performance. Finally, but certainly not least, this chapter will demonstrate the paper's uniqueness and how it intends to fill in the gaps left by previous studies on the subject.

Many studies have been conducted to investigate models that can be used to predict students' success based on their prior performance. For instance, in a study by Abu-Naser et al., the performance of sophomore engineering majors was predicted using an Artificial Neural Network (ANN) model based on a number of factors, such as high school scores, scores in subjects such as Math I, Math II, Electrical Circuit I, and Electronics I taken during the freshman year, number of credits passed, student cumulative grade point average of freshman year, types of high schools attended, and gender, among others. According to the study, the ANN model accurately predicted the academic performance of over 80% of prospective students (Abu-Naser, Samy S.; Zaqout, Ihab S.; Abu Ghosh, Mahmoud; Atallah, Rasha R.; Eman Alajrami; 2015). In another study conducted by Asif et al., admission data and examination scores of students in individual subjects from the first and second academic years were used to predict students' overall performance at the end of the degree. The study demonstrated that it is possible to predict graduation performance in the fourth year of university with a reasonable degree of accuracy and precision using only pre-university scores and examination scores from the first and second academic years, with no socioeconomic or demographic factors. The accuracy of the findings was determined to be 83.65 percent using Naive Bayes (Asif, R., Merceron, A., & Pathan, M. K.; 2014). A comparable conclusion is made in a study by Adejo et al. Using a quantitative research technique where data on 141 students enrolled at the University of the West of Scotland were pulled from the institution's databases and also acquired through a survey questionnaire, the study examined three data sources: student record systems, learning management systems, and surveys, and modeled them using three state-of-the-art data mining classifiers: decision trees, artificial neural networks, and support vector machines. The research demonstrates that combining numerous data sources with heterogeneous ensemble approaches is very effective and accurate in predicting student performance and identifying students at risk of attrition (Adejo, O. W., & Connolly, T.; 2018)

Furthermore, numerous studies conducted comparative research to see how efficient different modeling strategies are in predicting students' performance based on quantitative parameters. For example, in a study by Sokkhery et al., different strategies were studied and compared to find the best prediction model. For predicting student performance in mathematics, the study proposed a comparison study of statistical analytic methodologies, machine learning (ML) algorithms, and one of the deep learning architectures. Random Forest (RF) was found to outperform other models in predicting student performance in the research (Phauk Sokkhey; Takeo Okazaki; 2019). Ghorbani et al. conducted another study in which they compared multiple resampling methodologies for forecasting student performance using machine learning techniques. The study uses machine learning classifiers such as Random Forest, K-Nearest Neighbor, Artificial Neural Network, XG-boost, Support Vector Machine, Decision Tree, Logistic Regression, and Naive Bayes. The paper concludes by stating that the data from different methodologies indicate that models with fewer classes and nominal variables perform better. Additionally, it demonstrates that using balanced datasets improves classifier performance. Also, the paper shows that the Random Forest classifier outperforms all other models (Ghorbani, R., & Ghousi, R; 2020).

When it comes to growth models, however, there have been few studies that have looked at how successful this tool is in predicting student achievement. One stumbling block might be the shortage of longitudinal data, which contains detailed information on students' academic performance over time. However, in a study by Mirayama et al., growth models were used to predict students' long-term growth in mathematics. The study looked at how intelligence, motivation (perceived control, intrinsic motivation, and extrinsic motivation), and cognitive learning techniques (deep and surface strategies) all worked together to predict long-term improvement in students' mathematical achievement over a five-year period. The findings revealed that intelligence was highly linked to beginning levels of accomplishment, with motivation and cognitive methods accounting for further variance. On the other hand, intelligence showed no correlation with success increase over time, while motivation and learning methodologies were predictors of growth (Kou Murayama; Reinhard Pekrun; Stephanie Lichtenfeld; Rudolf vom Hofe; 2012). Furthermore, value-added models are widely regarded as one of the most promising solutions, not only for the reasons of accountability, but also for the goals of improving the educational system as a whole. A study conducted by Martin et al. focuses on this issue and proposes a method for estimating schools' added value under nonlinear growth models in which changes in performance follow a quadratic trajectory, analyzing the differences in results between the results obtained using linear growth models and those obtained using nonlinear growth models. The value-added in reading comprehension was estimated using three parallel cohorts from 153 primary and secondary schools in Madrid (Spain) and 6,755 students who were assessed at four different times during the academic years 2005–2006 and 2006–2007. The findings indicate that nonlinear growth models are more accurate and that the inclusion of the students' individual and family variables in the model results in more accurate value-added assessments for educational institutions and institutions of higher learning (Lopez-Martin, E., Kuosmanen, T., & Gaviria, J. L.; 2014)

In addition, numerous studies have been conducted with the objective of predicting students' performance based on external factors such as socioeconomic and demographic parameters. As an example, Ding et al. employed longitudinal multilevel modeling to see if there is a gender difference in mathematics performance on standardized as well as classroom assessments. The findings showed that, over time, both male and female students showed the same development trend in mathematical performance (as evaluated by standardized test scores), but that female students' mathematics grade-point average is much higher than male students' (Cody S. Ding; Kim Song; and Lloyd I Richardson; 2010). Another study conducted by Thiele et al. studied the relationships between school grades, school type, school performance, socioeconomic disadvantage, neighborhood involvement, gender, and academic accomplishment at a British university. The research found that students from the poorest neighborhoods did worse than those from more wealthy neighborhoods, that students of Asian and African descent performed worse than white students, and that female students outperformed their male peers. In contrast to earlier studies, however, students from low-performing schools were more likely to acquire the highest degree categories, even though school performance was positively related to admission grades. Additionally, despite their better starting grades, independent school students did worse in their final year than comprehensive school students. These variances indicate how trends detected at the national level may vary significantly between institutions (Thiele, T., Singleton, A., Pope, D., & Stanistreet, D.; 2016).

Additionally, several studies have been undertaken to determine how students would perform on standardized examinations. Crawford et al., for example, employed a curriculum-based measure (CBM) of reading aloud from narrative passages to predict success on statewide achievement exams in reading and mathematics. Scores on multiple-choice reading and arithmetic achievement exams were marginally linked with scores on rate measures conducted during the same year and a year earlier. The findings give preliminary evidence that timed oral readings may be used to predict students' success on statewide achievement examinations. (Lindy Crawford, Gerald Tindal & Steve.; 2010). Allensworth and colleagues did another study in which they analyzed the influence of high school graduation rates, high school GPA, and ACT scores on students' college graduation rates in order to determine which factors were most important. They found that students with similar high school grades or ACT scores graduate at significantly different rates depending on which high school they attended. However, the relationship between high school GPAs and college graduation is significant, consistent, and larger than the effects of individual schools or institutions of higher learning. However, when compared to high school graduation, the research found that the relationship between ACT scores and college graduation is weak and less significant than the relationship between high school graduation and college graduation, and that the slope of the relationship varies depending on which high school was attended. (Elaine M. Allensworth and Kallie Clark.; 2020).

This chapter begins by discussing earlier studies in which models were used to forecast students' success based on prior performance. Following that, the review analyzed prior research that examined the usefulness and relative advantages and disadvantages of several models for predicting student performance. Thirdly, the review offered insight into prior research that used growth models to predict student success. Fourth, the chapter discussed prior studies on the effect of external influences on student performance. Fifth, the chapter discussed prior research that examined the relationship between student performance and standardized test scores.

In light of the research reviewed before, this paper is distinctive in the following ways. To begin, this is the first publication to undertake a longitudinal examination of students' mathematics performance in the state of Rhode Island. Second, in contrast to earlier research that used artificial neural networks, generic clustering models, and the like, this paper will employ growth models to predict students' outcomes based on their performance in middle school mathematics. Third, in contrast to previous research that examined the effect of socioeconomic and other external factors on students' performance, this paper will examine whether there is a relationship between students' growth patterns, as determined by their previous performance, and their demographic factors, an approach that has not been attempted previously, particularly on a large scale given the longitudinal database.

CHAPTER 3: METHODOLOGY

The purpose of this study is to investigate the effectiveness of latent growth models in predicting students' outcomes in high school based on their performance in middle school mathematics. The study aims to answer the three questions: Can students be clustered based on their performance growth in middle school mathematics? Can the three demographic factors be used as indicators to predict students' growth performance clusters? Can students' growth performance clusters predict the students' academic outcome in high school?

Definitions

Growth performance clusters: Clusters that the study attempts to generate using latent growth models from students' growth performance in middle school mathematics.Students' performance in middle school mathematics: Data obtained from students' scores in their 6th, 7th, and 8th grades of the Rhode Island's Comprehensive Assessment System (RICAS),

The students' academic outcome in high school: The number of AP STEM courses passed in high school.

The three demographic predictors: Signs that are historically related to students' test performance; whether the student belongs to the historically-at-risk group (HAR), which includes Black, Latinos, and Native Americans; whether the student is eligible for a free lunch program (FLP); and whether the student is eligible for an individualized education plan (IEP),

Data collection

Data will be obtained from the Rhode Island Department of Education through Dr. Shane Tutwiler, an Associate Professor of Learning Foundations at the University of Rhode Island.

DataSpark, which is a service center at the University of Rhode Island, will remove all identifiable information from the data and clean the data to make it ready for analysis.

The data will include information about approx. 9,000 students. For each student, the following attributes will be provided: district, school, whether the student is in the high-risk status group (HR), whether the student is eligible for a free lunch program (FLP), whether the student is eligible for an individualized education plan (IEP), students' score in middle school mathematics obtained from the students' scores in the Rhode Island Comprehensive Assessment System (RICAS), students' score in high school mathematics obtained from the students' scores in RICAS, and the number of AP STEM courses passes.

Data optimization

Before running the analysis, a few steps will be taken to optimize the data. In addition to the standard procedures for deleting rows with missing data, the following procedure will be used: Only students whose demographic predictors have remained consistent throughout the middle school will be considered. This has no bearing on the final results of the data because only a few students exhibit variation in demographic predictors throughout middle school. This is because these demographic predictors are typically determined in elementary school and are rarely added or removed from students' profiles in middle school. The resulting data will be significantly easier to incorporate into a growth model.

Data Analysis

To answer the first question and cluster students based on their performance, the study will use the latent growth model in the *lavaan* package in R based on students' intercept (6thgrade score) and slope (score change over time). Next, k-mean will be used to generate clusters based on the growth model. The generated clusters will then be confirmed using SEM clustering.

To answer the second question, the likelihood ratio test will be used to establish a relationship between the three demographic predictors and students' growth performance clusters. Once this relationship is established, a multinomial logistic regression model will be used to find the likelihood that students with a specific demographic predictor are in one cluster over the other.

To answer the third question, the likelihood ratio test will be used to establish a relationship between students' growth performance clusters and academic outcome in high school, defined as the number of AP STEM classes passed in high school. Once this relationship Is established, the study will use the Poisson regression model in R to predict the students' academic outcome in high school, namely, the number of AP STEM courses passed in high school.

CHAPTER 4: RESULTS

Question 1

This section summarizes the results of question 1, “Can students be clustered based on their performance growth in middle school mathematics?”

The study uses the latent growth model found in the *lavaan* package in R to answer the first question and cluster students based on their performance. This model is based on students’ intercept (6th-grade score) and slope, calculated using the students’ intercept and slope (score change over time). After that, the k-mean algorithm is applied in order to form clusters based on the growth model. After that, SEM clustering is used to validate the clusters that were generated.

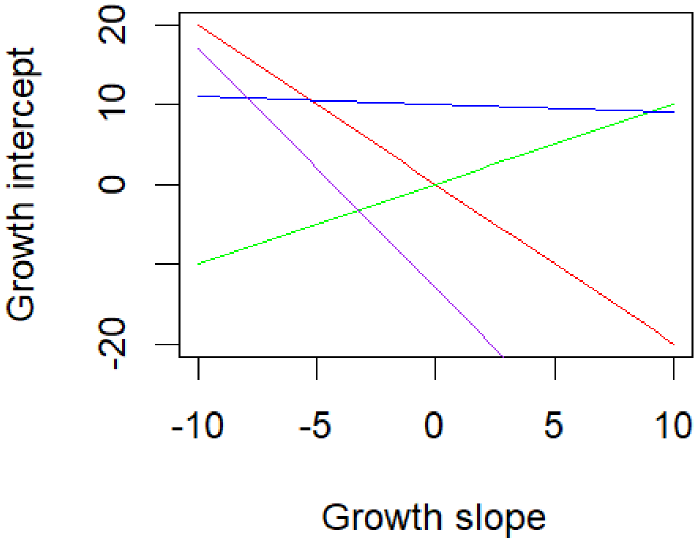


Figure 1: The four clusters based on the growth patterns of students

The study found that growth models can be effectively used to cluster students' growth performance based on their performance in middle school mathematics. Students were clustered in four clusters as per Figure 1. Below is a brief description and labeling of each of the clusters:

Cluster 1: Students with average growth performance (blue)

Cluster 2: Students with positive growth performance (green)

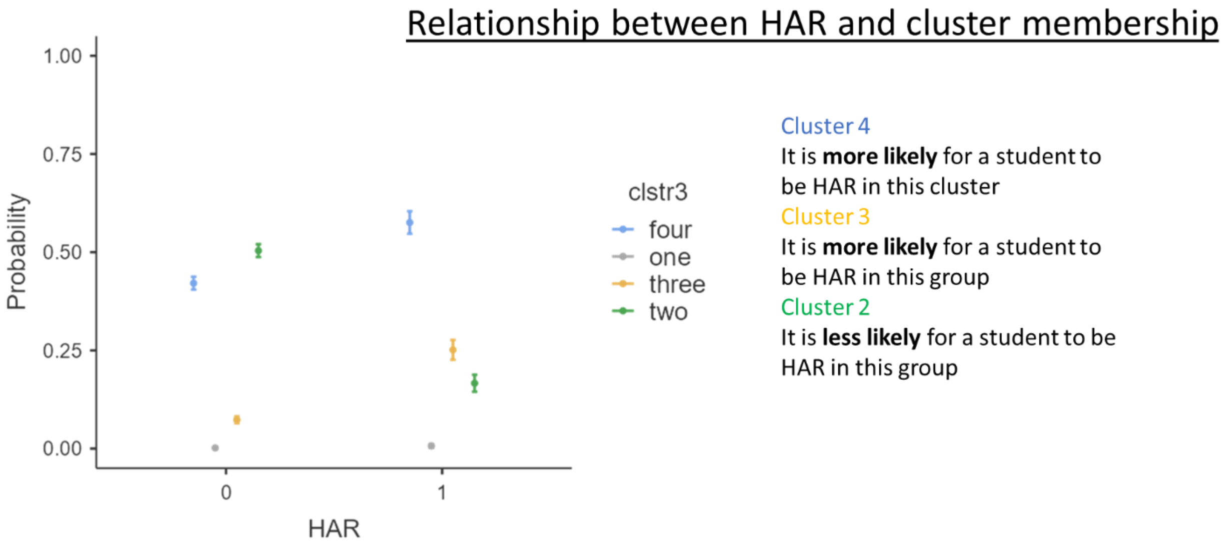
Cluster 3: Students with negative growth performance (orange)

Cluster 4: Students with very negative growth performance (purple)

Question 2

This section summarizes the results of question 2, "Can the three demographic factors be used as indicators to predict students' growth performance clusters?"

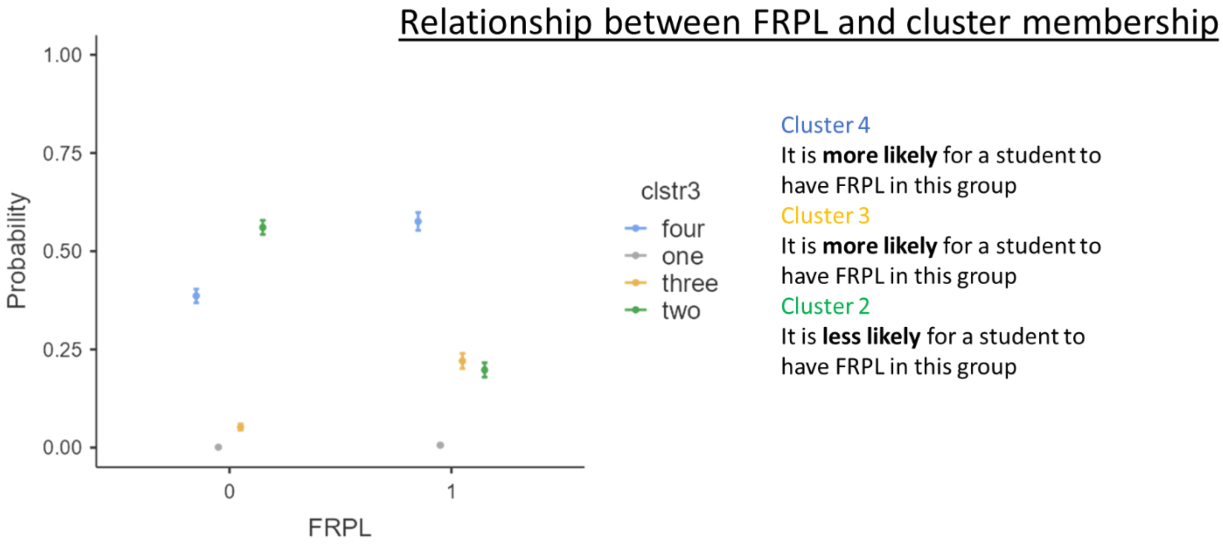
In order to provide an answer to the second question, the likelihood ratio test is applied to determine whether or not there is a connection between the three demographic predictors and the growth performance clusters of the students. After this connection has been made, a multinomial logistic regression model is used in order to determine the likelihood that students who share a particular demographic predictor are members of one cluster rather than the other.-


Relationship between HAR and cluster membership

Cluster 4
It is **more likely** for a student to be HAR in this cluster
Cluster 3
It is **more likely** for a student to be HAR in this group
Cluster 2
It is **less likely** for a student to be HAR in this group

Figure 2: Relationship between students who belong to the historically-at-risk group and cluster membership

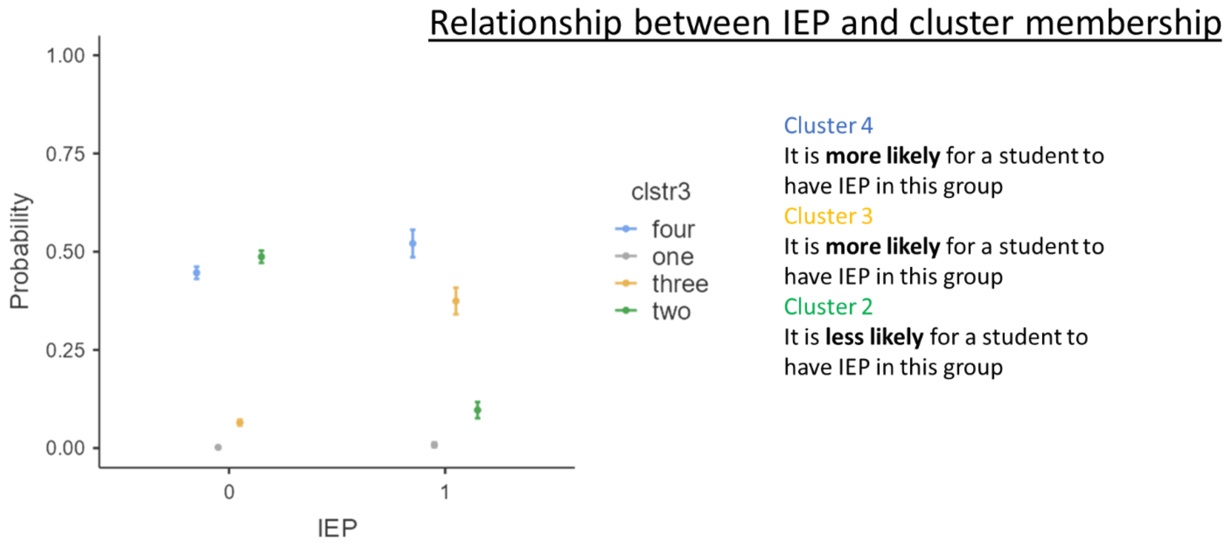
Figure 2 above summarizes the relationship between HAR students (students who belong to the historically-at-risk group) and cluster membership. The Figure demonstrates that the likelihood of a student being HAR increases in clusters 3 and 4, but decreases in cluster 2, where the likelihood of a student being HAR is lowest.


Relationship between FRPL and cluster membership

Cluster 4
It is **more likely** for a student to have FRPL in this group
Cluster 3
It is **more likely** for a student to have FRPL in this group
Cluster 2
It is **less likely** for a student to have FRPL in this group

Figure 3: Figure 4: Relationship between students eligible for a free lunch program and cluster membership

Likewise, Figure 3 above summarizes the relationship between FLP students (students eligible for a free lunch program) and cluster membership. The Figure demonstrates that the likelihood of a student being HAR increases in clusters 3 and 4, but decreases in cluster 2, where the likelihood of a student being HAR is lowest.



Relationship between IEP and cluster membership

Cluster 4
It is **more likely** for a student to have IEP in this group
Cluster 3
It is **more likely** for a student to have IEP in this group
Cluster 2
It is **less likely** for a student to have IEP in this group

Figure 5: Figure 6: Relationship between students eligible for an individualized education plan and cluster membership

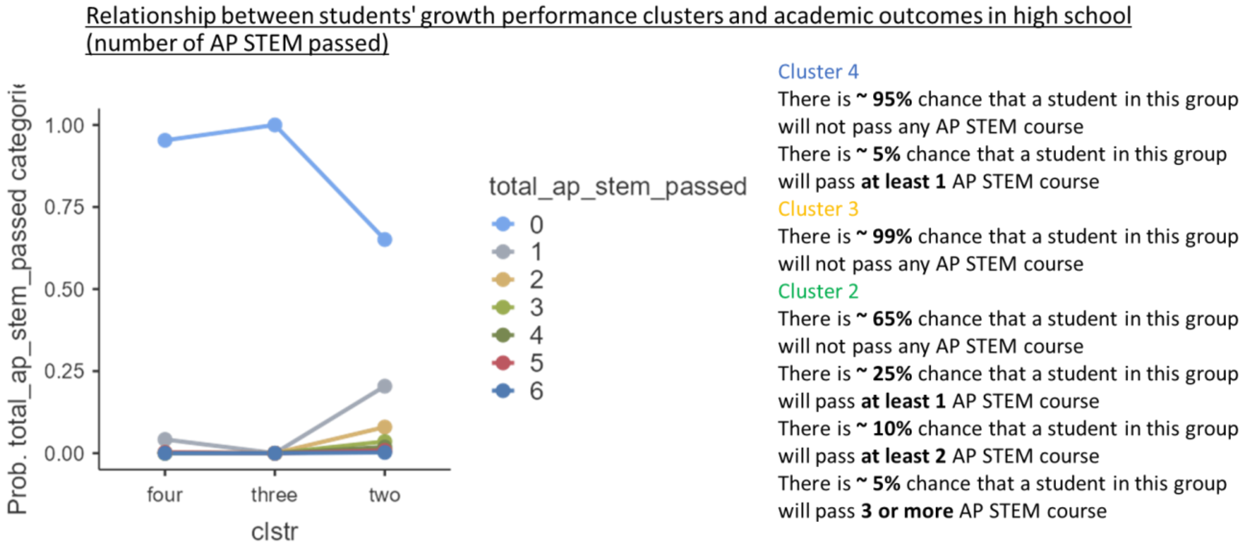
Finally, Figure 4 above summarizes the relationship between IEP students (students eligible for an individualized education plan) and cluster membership. The Figure demonstrates that the likelihood of a student being HAR increases in clusters 3 and 4, but decreases in cluster 2, where the likelihood of a student being HAR is lowest.

According to the findings and Figures 2, 3, and 4 above, the clusters that were generated are consistent with the expected patterns that have been demonstrated historically. It is, therefore, possible to draw the conclusion that it is more likely for students to belong to clusters that show negative growth performance because they have the three demographic predictors.

Question 3

This section summarizes the results of question 3, “Can students’ growth performance clusters predict the students’ academic outcome in high school?”

In order to provide an answer to the third question, the likelihood ratio test is applied in order to establish a connection between the growth performance clusters of students and their academic outcome in high school, which is defined as the number of Advanced Placement Science, Technology, Engineering, and Math (AP STEM) classes that were successfully completed in high school. Once this relationship has been established, the research uses the Poisson regression model in R to predict the students’ academic outcome in high school, which is defined in this study as the number of AP STEM classes that the students passed.



Relationship between students' growth performance clusters and academic outcomes in high school (number of AP STEM passed)

Cluster 4

There is **~ 95%** chance that a student in this group will not pass any AP STEM course

There is **~ 5%** chance that a student in this group will pass **at least 1** AP STEM course

Cluster 3

There is **~ 99%** chance that a student in this group will not pass any AP STEM course

Cluster 2

There is **~ 65%** chance that a student in this group will not pass any AP STEM course

There is **~ 25%** chance that a student in this group will pass **at least 1** AP STEM course

There is **~ 10%** chance that a student in this group will pass **at least 2** AP STEM course

There is **~ 5%** chance that a student in this group will pass **3 or more** AP STEM course

Figure 7: Relationship between students‘ growth performance clusters and academic outcome in high school (number of AP STEM passed)

Figure 5 above shows the relationship between students’ growth performance clusters and academic outcome in high school (number of AP STEM passed)

It shows that in Cluster 4, which has students with very negative growth performance, there is ~ 95% chance that a student in this group will not pass any AP STEM course; and that there is ~ 5% chance that a student in this group will pass at least 1 AP STEM course

In addition, Figure 5 shows that in Cluster 3, which has students with negative growth performance, there is ~ 99% chance that a student in this group will not pass any AP STEM course.

Furthermore, Figure 5 shows that in Cluster 2, which has students with positive growth performance, there is ~ 65% chance that a student in this group will not pass any AP STEM course; there is ~ 25% chance that a student in this group will pass at least 1 AP STEM course; that there is ~ 10% chance that a student in this group will pass at least 2 AP STEM course; and that there is ~ 5% chance that a student in this group will pass 3 or more AP STEM course.

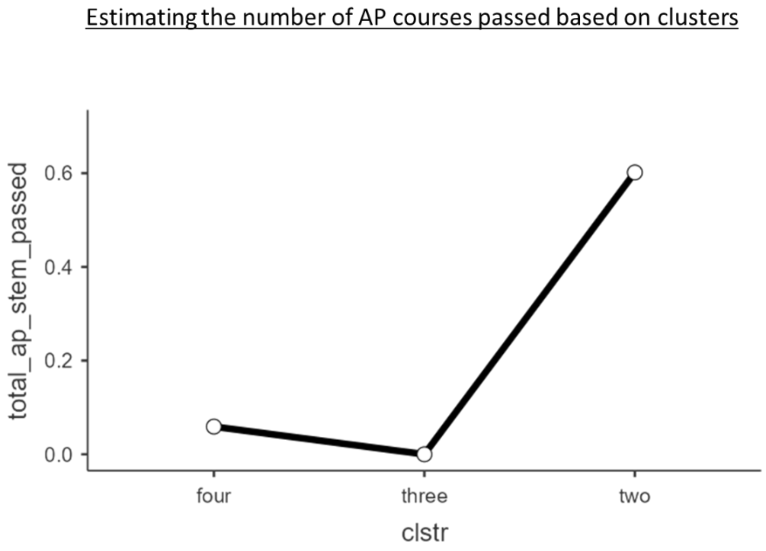


Figure 8: Estimating the number of AP courses based on clusters

Finally, Figure 6 above shows a general estimation of the number of AP STEM courses passed. It shows that students in cluster 2 pass more ap stem courses than clusters 3 and 4, which is consistent with the cluster designation provided in Question 1.

Quantitatively, the estimated number of AP courses passed for students in clusters 4, 3, 2 is 0.8416757, 1.039676, and 2.600783, respectively.

Conclusion

In summary, the study came to the conclusion that the growth of students can be broken down into clusters. It also showed that membership in a cluster was associated with key demographic predictors that were consistent with the expected trends that had been shown historically. In addition, the study found that the students' performance in high school can be predicted based on their membership in a growth cluster.

These findings can be used to determine which students who belong to negative growth performance clusters need to be helped by taking appropriate intervention measures.

CHAPTER 5: SUMMARY

Despite the abundance of data, many educational institutions lack an analytical foundation to use large data to make effective, evidence-based decisions that improve students' education quality and help educators provide necessary intervention to ensure all students have equal opportunity. This research aims to help the Rhode Island Department of Education improve student education.

This study is unique because it uses latent growth models to cluster students' performance based on their middle school math performance, investigates the extent to which demographic factors can predict these performance clusters, and predicts students' high school academic outcome based on these performance clusters. It is also the first attempt to cluster students in the state of Rhode Island based on their growth pattern that is obtained by utilizing latent growth models from students' performance in middle school mathematics

This study investigates the utility of latent growth models in clustering middle school math students to predict the high school academic outcome. Middle school math performance is based on 6th, 7th, and 8th-grade RICAS scores (RICAS). Students' outcomes refer to high school performance, specifically passing AP STEM courses. The study also examines whether demographic factors predict student performance clusters. High-risk status, free lunch eligibility, and/or an individualized educational plan (IEP). The study tests and predicts. Latent growth models will be used to analyze data and accept or reject hypotheses. The analysis will result in a model that can predict the academic outcome based on middle school math performance clusters and the three demographic factors (HR, FLP, and IEP).

The study found that growth models can be utilized effectively to cluster students' growth performance based on their middle school mathematics performance, which clustered students into four groups. The study also found that students with demographic factors are more likely to be found in clusters with low performance than students without demographic factors. The study concluded that latent growth models can predict high school performance as measured by the number of AP STEM courses taken and passed.

Since the number of passed AP STEM courses determines the scope of the study, the study makes no mention of dual enrollment, honors courses, extracurricular activities, or college admission. There is no mention of other indicating factors that could be considered performance factors, such as non-mathematics performance or classroom interaction, in this study. To maximize the efficacy of the research, only middle school students with constant demographic factors were included. This resulted in a negligible decline in the number of students, which had no bearing on the study's findings. Other factors, such as the school district or high school, may also influence the academic outcome in high school. The study used multinomial regression to account for confounding factors.

REFERENCES

Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education*, *10*(1), 61–75. https://doi.org/10.1108/jarhe-09-2017-0113

Allensworth, E. M., & Clark, K. (2020). High School GPAS and ACT scores as predictors of college completion: Examining assumptions about consistency across high schools. *Educational Researcher, 49*(3), 198–211. https://doi.org/10.3102/0013189x20902110

Asif, R., Merceron, A., & Pathan, M. K. (2014). Predicting student academic performance at Degree Level: A Case Study. *International Journal of Intelligent Systems and Applications*, *7*(1), 49–61. https://doi.org/10.5815/ijisa.2015.01.05

DING, C. O. D. Y. S., SONG, K. I. M., & RICHARDSON, L. L. O. Y. D. I. (2006). Do mathematical gender differences continue? A longitudinal study of gender difference and excellence in mathematics performance in the U.S. *Educational Studies, 40*(3), 279–295. https://doi.org/10.1080/00131940701301952

Downy, R., Rubin, D., Cheng, J., & Bernstein, J. (n.d.). *Performance of automated scoring for Childrens Oral Reading - ACL anthology*. ACL Anthology. Retrieved August 8, 2022, from https://aclanthology.org/W11-1406.pdf

Ghorbani, R., & Ghousi, R. (2020). Comparing different resampling methods in predicting students' performance using Machine Learning Techniques. *IEEE Access*, *8*, 67899–67911. https://doi.org/10.1109/access.2020.2986809

KABACOFF, R. O. B. E. R. T. I. (2021). *R in action*. MANNING PUBLICATIONS.

Lopez-Martin, E., Kuosmanen, T., & Gaviria, J. L. (2014). Linear and nonlinear growth models for value-added assessment: An application to Spanish primary and secondary schools' progress in reading comprehension. *Educational Assessment, Evaluation and Accountability*, *26*(4), 361–391. https://doi.org/10.1007/s11092-014-9194-1

Murayama, K., Pekrun, R., Lichtenfeld, S., & vom Hofe, R. (2012). Predicting long-term growth instudents' mathematics achievement: The unique contributions of motivation and cognitive strategies. *Child Development*, *84*(4), 1475–1490. https://doi.org/10.1111/cdev.12036

Naser, S. A., Zaqout, I., Ghosh, M. A., Atallah, R., & Alajrami, E. (2015). Predicting student performance using Artificial Neural Network: In the Faculty of Engineering and Information Technology. *International Journal of Hybrid Information Technology*, *8*(2), 221–228. https://doi.org/10.14257/ijhit.2015.8.2.20

Sokkhey, P., & Okazaki, T. (2020). Study on dominant factor for academic performance prediction using feature selection methods. *International Journal of Advanced Computer Science and Applications*, *11*(8). https://doi.org/10.14569/ijacsa.2020.0110862

Thiele, T., Singleton, A., Pope, D., & Stanistreet, D. (2014). Predictingstudents' academic performance based on school and Socio-demographic characteristics. *Studies in Higher Education*, *41*(8), 1424–1446. https://doi.org/10.1080/03075079.2014.974528

BIBLIOGRAPHY

Sam Habach

Candidate for the Degree of

Master of Science Mathematics

Thesis: PREDICTING STUDENTS’ PERFORMANCE USING GROWTH MODELS

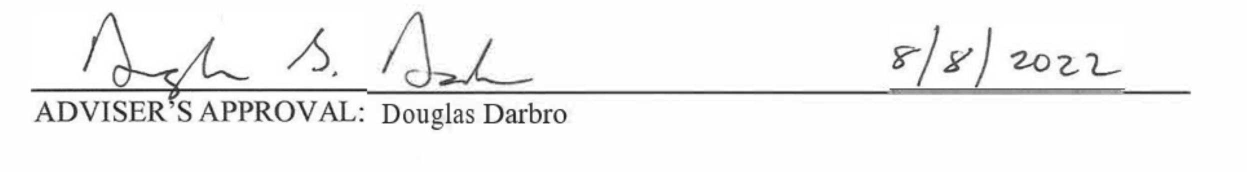
Major Field: Mathematics

Biographical:

Personal Data:

Education: (prior degrees): Bachelor of Sscience in Physics and Master of Science in Data Analytics

Completed the requirements for the Master of Science in Mathematics, Portsmouth, Ohio in August 2022.



8/8/2022

ADVISER’S APPROVAL: Douglas Darbro