PREDICTORS OF EARLY POSTSECONDARY STEM PERSISTENCE OF HIGH-ACHIEVING STUDENTS: AN EXPLANATORY STUDY USING MACHINE LEARNING TECHNIQUES

by

Nesibe Karakis

A Dissertation

Submitted to the Faculty of Purdue University
In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Educational Studies

West Lafayette, Indiana

August 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Nielsen Pereira, Chair

Department of Educational Studies

Dr. Hua-Hua Chang

Department of Educational Studies

Dr. Marcia Gentry

Department of Educational Studies

Dr. Kristen Seward

Department of Educational Studies

Dr. Anne Traynor

Department of Educational Studies

Approved by:
Dr. Ayse Ciftci

Dedicated to my loving, encouraging, and supporting family

ACKNOWLEDGMENTS

I have received a lot of support and help throughout my dissertation as well as my education at Purdue University. I would like to thank the great individuals who have helped me along this journey. I am blessed and continually grateful to have you all in my life. I would also like to thank the Turkish Ministry of National Education for supporting my work in the United States.

I would like to express my appreciation to my advisor, Dr. Nielsen Pereira. I am deeply indebted to you. Thank you for accepting me as your mentee and for kindly supporting me throughout my graduate education. I knew that you were always there for me when I needed your help, guidance, and support. You have always been kind, helpful, and understanding and supported my research interests. I would not be where I am today without your help and support. Thank you for being such a great advisor.

I would also like to extend my deepest gratitude to Dr. Marcia Gentry for supporting me throughout my education and serving on my committee. I was able to attend great conferences and lucky to take enriching courses from you. I also thank you for your valuable feedback during the development of my studies. You are a great leader and a great professor in the field. Thank you for your continuous support.

I am extremely grateful to Dr. Hua-Hua Chang for being my committee member and providing great feedback during the development of my dissertation. It was an honor to have you serve on my committee. The completion of my dissertation would not be possible without your support and guidance. Thank you so much.

I also would like to thank Dr. Anne Traynor for serving on my committee and helping me throughout my education. I took many great courses with you and was so lucky to have you serve

on my committee. You have always been very helpful and kind. Thank you for your valuable advice, support, and feedback.

Next, I would like to thank Dr. Kristen Seward for taking your time out of your busy schedule and serving on my committee. We shared many great experiences, and it was an honor to know and work with you. Thank you for being there when I needed help.

I also would like to thank Dr. Ala Samarapungavan. You have been a great supervisor. I really enjoyed working with you. Your work always inspired me and taught me a lot. Being a member of your research group and working under your guidance was such an honor. Thank you so much.

I also would like to thank my great GER^{2}I family, being a part of this fantastic group was indeed a blessing. I thank Drs. Hodges, Yi, Tay, Desmet, Karami, Ghahremani, Gray, Green, and Lee for always helping me and being great friends. You all are great professionals in the field. I am happy to have spent a great deal of time together during our GER^{2}I programs and graduate education. I would also like to thank Abdullah, Alissa, Andres, Aakash, and Bekir for being great colleagues and collaborators. You all are amazing and doing great work.

I would also like to thank my dear friend Merve. We started this journey together. I cannot thank you enough for your support and priceless friendship. You are a critical thinker with great ideas, and you truly inspire me every day. I am so happy that you are in my life. Thank you for being a wonderful friend. Also, I would like to thank all of my friends who I could not individually mention but have been a great part of my life.

Last but not least, I would like to thank my loving family. I would like to express my sincere thanks to my parents as well as my sisters Fatma, Ayşe Nur, and Hatice Kübra. You are the light of my life. Words cannot express how grateful I am for having you. You always believed

in me and encouraged me to do my best. I appreciate your love, patience, and understanding as I pursued my degree. Without your support, it would not be possible. I am so lucky to have you, and I always know that you will be there for me. Thank you for being in my life. I also thank my brother in-laws İsmail and Cahit. Even though you joined our family later, I felt as if you were always a part of our family. Thank you for being my big brothers. I also thank my nieces Hatice Erva, Zeynep Gülce, Gökçe Vefa, and my nephew Emre Mehmet for always cherishing me and for making me the happiest aunt in the world. Having you is such an incredible joy.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This study investigated high-achieving and non-high-achieving students' persistence in STEM fields using nationally representative data from the High School Longitudinal Study of 2009 for the years 2009, 2012, 2013, 2013-2014, and 2016. The results indicated that approximately 70% of high-achieving and non-high-achieving students continued their initial STEM degrees within 3 years of college enrollment. The study revealed that the most important predictors of STEM persistence were: math proficiency level, school belonging, school engagement, school motivation, school problems, science self-efficacy, credits earned in computer sciences, GPA in STEM courses, credits earned in STEM courses, and credits earned in Advanced Placement/International Baccalaureate (AP/IB) courses. Based on the results, math proficiency was the most important variable in the study for both high-achieving and non-high-achieving students. Even though credits earned in AP/IB combined were among the most important variables, they were two times more important for high-achieving students (6.86% vs. 3.37%). Regarding demographic information related variables, socioeconomic status was the most important variable among gender, ethnicity, and urbanicity in models predicting STEM persistence and had higher importance for non-high-achieving students. Furthermore, Hispanic students' proportion of persistence differed from other underrepresented populations' persistence. Non-high-achieving Hispanic students had the highest persistence rate, similar to well-represented populations (i.e., White, Asian). Machine learning methods used in the study including random forest and artificial neural network provided good accuracy for both achievement groups. Random forest accuracy was over 82% with the Synthetic Minority Over-Sampling Technique (SMOTE) dataset, while artificial neural network accuracy was over 92%.

CHAPTER 1 INTRODUCTION

Society needs large numbers of students in Science, Technology, Engineering, and Mathematics (STEM) fields as these areas support innovation, technical development, global competitiveness, and economic growth (Maltese et al., 2014). Based on the National Science and Engineering Indicators (National Science Board [NSB], 2018) report, STEM education is a priority globally due to current higher demand and also faster growing job opportunities in STEM than non-STEM fields. The report revealed, however, that the percentage of students who completed Science and Engineering degrees in India, China, the European Union, and the United States was 25%, 22%, 12%, and 10%, respectively, from 2000 to 2014. In particular, China's STEM degree attainment increased by 350% from 2000 to 2018 (NSB, 2018), while in the United States enrollment rates in STEM areas declined. Based on National Center for Education Statistics (NCES) report, only 48% of students who initially declared STEM majors continue their education in STEM fields within three years of college enrollment in the United States (Chen & Soldner, 2013). On the other hand, another report by NCES (2017) regarding first-time postsecondary students' persistence after three years of initial enrollment revealed that 70% of students persisted at their institutions in general. To avoid falling behind other countries, the United States needs to significantly improve support of STEM education and higher education degree attainment.

Researchers have recently promoted the importance of supporting excellence in STEM fields in the United States and encouraging gifted and talented students to pursue studies in these areas (Heilbronner, 2011; Steenbergen-Hu & Olszewski-Kubilius, 2017). Nonetheless, despite their high potential, a small proportion of gifted and talented students enter into STEM fields (Holmes et al., 2018). Furthermore, gifted and talented students do have low STEM persistence in college (Heilbronner, 2011; Steenbergen-Hu & Olszewski-Kubilius, 2017) in STEM areas. Dweck

(2002) pointed out that students' view of learning (e.g., fixed or growth mindset) affects their persistence. Heilbronner (2011) suggested that many students who drop out of STEM fields could be successful if they persisted and believed in their abilities. However, there is not sufficient research on the early postsecondary STEM persistence of high-achieving students. Many researchers (e.g., Green & Sanderson, 2018; Mendez et al., 2008) focused on the general student population. Thus, continued research is needed to understand the STEM persistence of high-achieving students, as they are more likely to have the potential to fulfill the need for graduates with STEM expertise.

According to Olszewski-Kubilius (2006), school experiences (e.g., math and science courses) might not be challenging or motivating students to pursue careers in STEM fields. She suggested that additional experiences outside of school (e.g., summer and weekend camps) can engage students in STEM learning and ultimately in pursuing STEM degrees. Furthermore, Green and Sanderson (2018) found that students' educational experiences were not a significant predictor of persistence among students, whether or not they were interested in STEM, suggesting the importance of out-of-school programs.

Additionally, the effects of educational experiences on STEM persistence are even lower for high-achieving students who are from underrepresented populations. Previous studies have shown that students from racially, culturally, economically, and linguistically diverse populations are less likely to attend a postsecondary institution and even less to attain a degree (Ashford et al., 2016; Diemer & Li, 2012), with the exception of Asian students, who are more likely to attain STEM degrees than their Black, Latinx, and Native American counterparts (Mendez et al., 2008). Many researchers have documented the barriers that underrepresented students face including racism; sexism; low science and math grades in high school; and insufficient STEM career

information; and they are less likely to pursue careers in STEM (Anderson, 2016; Assouline et al., 2017; Turner et al., 2019). Wai et al. (2010) investigated the predictors of STEM achievement in college based on students' advanced or enriched academic experiences in high school and found that students who had taken more advanced STEM courses and who had more opportunities in STEM were high achievers. However, many underrepresented students take a limited number of STEM courses in high school compared to their Asian American and White American peers leading to low STEM persistence in college (Ashford et al., 2016; Maltese & Tai, 2011).

Given the importance of understanding factors that affect students' persistence in STEM fields, researchers have focused on demographic, cognitive, and non-cognitive variables (Aryee, 2017; Mendez et al., 2008). Quantitative analytic methods have included logistic regression (Mendez et al., 2008; Tyson et al., 2007; Watkins & Mazur, 2013; Zhang et al., 2004), discriminant function analysis (Achter et al., 1999), hazard/survivor models (Chimka et al., 2007; Min et al., 2011), repeated measures ANOVA (Eris et al., 2010), and multinomial probit models (Chen & Soldner, 2013). Furthermore, with the development of machine learning, researchers have been widely used machine learning methods to examine factors affecting student persistence (Adejo & Connolly, 2018; Delen, 2010; Dissanayake et al., 2016; Kondo et al., 2017; Pereira et al., 2017). However, there is a need to understand high-achieving and non-high-achieving students' persistence in STEM areas as there is limited research in the field.

Compared to the traditional statistical methods, machine learning methods can be used to examine more complex relationships (Mendez et al., 2008). These methods provide better predictive results and have no limitations such as normality and independence. Furthermore, they can deal with missing data and nonlinear relationships, especially when working with large datasets (Thammasiri et al., 2014). In this study, I used machine learning methods to identify

variables affecting early post-secondary STEM persistence to better understand how they can contribute to the literature on early prediction of postsecondary STEM persistence.

Significance of Study

The findings of this study add to the literature on STEM persistence. In particular, using machine learning methods have the potential to identify variables that may be related to early postsecondary STEM persistence but that may not be identified with other standard statistical methods. The prediction of whether or not a student persist in an initial STEM major is not an easy task due to the complex relationships between and among variables (Mendez et al., 2008). Finding the predictors of persistence using traditional methods might not be sufficient to examine all the variables associated with persistence due to the limited number of variables that can be included in the models.

Researchers have investigated different variables and their association with students' STEM persistence, such as high school GPA (Mendez et al., 2008; Nicholls et al., 2007; Zhang et al. 2004), SAT math scores (Chimka et al., 2007; French et al., 2005; Min et al., 2011; Nicholls et al., 2007; Watkins & Mazur, 2013; Zhang et al., 2004), gender (Chimka et al., 2007; Min et al., 2011; Nicholls et al., 2010), number of STEM courses taken (Chen & Soldner, 2013; Mendez et al., 2008; Nicholls et al., 2010; Tyson et al., 2007; Wang, 2013), and experiences in high school math and science curriculum (Adelman, 1998). However, the effects of non-cognitive factors such as interest, outcome expectations, and self-efficacy (Aryee, 2017) have rarely been investigated. Additionally, studies of high-achieving students' STEM persistence are scarce (Anderson, 2016; Yi, 2018).

Specifically, focusing on demographics, cognitive, and non-cognitive variables that have the potential to affect persistence in STEM in the same study might provide better results. Also,

variables associated with persistence might not be found by other standard statistical methods. In this study, feature selection techniques are used to select the best predictor variables of STEM persistence. The selected machine learning methods in this study might be used in other settings, especially in similarly designed longitudinal studies, to predict students' early postsecondary STEM persistence. The results may inform efforts to support students' persistence in STEM fields and identify students who are more likely to drop out of STEM fields within three years of college entrance.

Purpose of the Study

In this study, I investigated factors providing early prediction of postsecondary STEM persistence using machine learning methods. Traditional statistical methods commonly used to examine relationships between variables and their effects on persistence (Hodges & Mohan, 2019; Kučak et al., 2018; Mason et al., 2018). Machine learning techniques (e.g., random forest and artificial neural network) can also be implemented to examine more complex relationships between and among the variables. These new techniques can contribute to the current STEM persistence literature. To examine the predictors of early postsecondary STEM persistence of high-achieving as well as non-high-achieving students from diverse populations, the following research questions were addressed in this study:

1. What percentages of high-achieving and non-high-achieving students persist in postsecondary STEM majors?
2. What variables affect high-achieving and non-high-achieving students' persistence in postsecondary STEM majors?
3. Which variables most significantly influence the early postsecondary persistence of high-achieving and non-high-achieving students?

4. To what extent do high-achieving and non-high-achieving students' demographics (e.g., gender, ethnicity, socioeconomic status) affect their early postsecondary STEM persistence?
5. Which machine learning techniques can be used to identify variables influencing the early postsecondary persistence of high-achieving and non-high-achieving students in terms of classification models?

To answer these questions, I used nationally representative data, the High School Longitudinal Study of 2009 (HSLS: 2009), collected by the National Center for Education Statistics (NCES). These data can be used to investigate students' transitions from high school to college as well as to the job market. Using the data, I investigated demographics, cognitive factors, and non-cognitive factors of the early postsecondary STEM persistence of high-achieving students. To select variables associated with early postsecondary STEM persistence, I used feature selection techniques, which are essential preprocessing steps used in machine learning to find the predictor variables that increase the model's performance (Saeys et al., 2008). Thus, these techniques are used to select the best variables to create a model. For instance, in the literature, parents' education level has been shown to influence student persistence (Nicholls et al., 2010). In this study, instead of selecting only the highest education level of one parent as a variable, I included all the variables associated with parent education such as mother's, father's, female guardians', male guardians', and parents' highest levels of education in feature selection. Using feature selection techniques helped me to select the best variables for the predictive model of students' persistence in STEM. In particular, feature selection is an essential task in machine learning to create an effective predictive modeling system. After choosing the variables, I used

Random Forest and Artificial Neural Networks methods in this study. Then, using selected variables, based on the acceptable machine learning prediction performance, random forest feature importance technique was used to examine which variables were more important for predicting early postsecondary STEM persistence.

CHAPTER 2 LITERATURE REVIEW

Although previous studies have highlighted the importance of a qualified workforce in STEM areas (Ashford et al., 2016; Holmes et al., 2018), student persistence in STEM education still appears to be problematic (Heilbronner, 2011; Simon et al., 2015; Steenbergen-Hu & Olszewski-Kubilius, 2017). The enrollment and completion of postsecondary STEM degrees have been declining in the U.S (Holmes et al., 2018; Steenbergen-Hu & Olszewski-Kubilius, 2017) regardless of gender (Mendez et al., 2008). Although the United States attracts many students in STEM fields from all over the world, between 2000 and 2014, international student enrollment in U.S. postsecondary institutions decreased from 25% to 19% in these fields (NSB, 2018). As a result, due to low enrollment and completion rates, there is a need to support STEM areas in the United States.

In the United States, a large proportion of college students who begin with STEM leave these fields prior to their postsecondary degree attainment because they either switch to non-STEM areas or drop out of college (Green & Sanderson, 2018). According to Chen and Soldner's (2013) report based on nationally representative data, 48% of students who entered postsecondary STEM fields did not remain until graduation. Specifically, 28% switched to non-STEM fields, and 20% of them left the institution (Chen & Soldner, 2013). This situation creates a "leak" in the STEM pipeline. According to Holmes et al. (2018), the problems of decreasing enrollments and a lack of persistence have to be solved to meet the growing needs for a qualified workforce in STEM areas. Thus, it is crucial to encourage students to pursue and persist in STEM majors. According to Chen (2009), STEM persistence is defined in terms of students who continue their initial STEM degrees within 3 years of college enrollment. One way to increase the number of students who persist to graduation in STEM fields is to better understand and support them during their undergraduate

education. Thus, every type of decline in degree completion in STEM including switching to non-STEM majors and dropping out of school should be reduced (Aryee, 2017; Simon et al., 2015).

Given the critical need, researchers have studied the variables that predict STEM persistence to identify factors at both high school and postsecondary levels likely to yield persistence. Steenbergen-Hu and Olszewski-Kubilius (2017) found the longitudinal process of STEM postsecondary degree attainment is linked with students' secondary education. STEM persistence is a longitudinal process, and secondary education has a significant impact on the completion of STEM degrees in college (Steenbergen-Hu & Olszewski-Kubilius, 2017). Additionally, Shaw and Barbuti (2010) studied the patterns of persistence of 54,336 third year college students and found that advanced placement exams, performance in science and math courses, positive science efficacy beliefs, and aiming for a doctorate were factors positively associated with STEM persistence in college. Similarly, Green and Sanderson (2018) found that the number of math and science courses taken in high school was linked with persistence in STEM and stated that there is little evidence that educational experiences affect persistence in college. With regard to who were confident in their high school science and math skills were more likely to persist in STEM areas at postsecondary levels (Eris et al., 2010). To augment these prior findings, more research is needed to determine the most effective predictors of student persistence in postsecondary STEM education so that improvements can be made to support the persistence of future STEM students.

Machine Learning

Machine learning is an important technology that enables computers to learn from past data and apply what is learned to new situations without being explicitly programmed (Samuel, 1959). Machine learning has existed since the mid-1900s; however, it has not been well known and used

in research, especially in education, until recent years (Kučak et al. 2018; Hodges & Mohan, 2019). According to Richert and Coelho (2013), machine learning is not an entirely new field because of the use of techniques and knowledge similar to statistics. Machine learning is instead used to understand underlying patterns and relationships in the data to make predictions (Delen, 2010). With the improvement of computer technologies, machine learning techniques have gained increasing popularity due to their ability to yield fast and accurate results (Kučak et al. 2018). Recently, machine learning algorithms have been used in several areas such as computer vision, natural language processing, and automated driving; however, applications of machine learning in education remain limited (Richert & Coelho, 2013).

Some researchers have proposed that machine learning methods could be used in education to support teachers, predict student performance, and personalize student learning (Kučak et al., 2018; Maseleno et al., 2018). For instance, Pavleković et al. (2011) demonstrated that a machine learning method known as a neural network, could identify fourth grade students who were mathematically talented. Okubo et al. (2017) showed that predicting students' final grades with machine learning methods was more accurate and took less time than standardized statistical methods. Such efficiency is possible because machine learning uses various techniques and computational methods to increase its performance in making predictions (Kučak et al., 2018).

According to Richert and Coelho (2013), similar to other statistical techniques, machine learning requires researchers to follow specific steps:

1. Read and clean the data.
2. Explore and understand the given data.
3. Analyze how best to present the data to the learning algorithm.

4. Choose the right model and the learning algorithm.
5. Measure performance accurately.

In machine learning, when working with data, researchers generally split data into two parts as training and test sets. Experts often use the terms “validation” and “test” sets interchangeably. The majority of data in machine learning are used to train the model so that it can perform well and provide accurate results. The rest of the data are used to confirm an unbiased evaluation of the final model that fits the training data.

One of the most important machine learning method categories is supervised learning, in which the target (dependent) variable is labeled, and the input data are used to find the target value (Müller & Guido, 2016). To examine the target value, supervised learning uses its initial output against the target output so that weights and coefficients can be adjusted based on these two outputs (Alpaydin, 2004; Hodges & Mohan, 2019). The supervised learning algorithms are complex and yield accurate results.

Hodges and Mohan (2019) and Kučak et al. (2018) have asserted that understanding and using the framework and techniques of machine learning in education could provide benefits to the field such as generating more accurate results, solving complex problems, and making predictions.

Predictor Variables of STEM Persistence

Interest

Interest is one of the most studied latent constructs and a significant predictor of STEM persistence (Aryee; 2017; Heilbronner 2011, 2013; Lubinski et al., 2001; Steenbergen-Hu & Olszewski-Kubilius, 2017). Interest indicates the degree to which individuals prefer specific tasks

over others (Lent et al., 1994, 2000). According to Aryee (2017), students who have a strong interest in pursuing a STEM degree are more likely to persist and complete initially declared STEM majors. Researchers have shown that students tend to lack interest in STEM, which leads to declines in enrollment rates in STEM areas (Heilbronner, 2011, 2013; Holmes et al., 2018).

In a longitudinal study of talented students Heilbronner (2011) found that 74.2% of those students who had STEM interests in high school completed STEM degrees. This result suggests the need for further investigation of the role of interest in students' success in STEM majors at the college level. However, Nugent et al. (2015) argued that interest does not directly correlate with STEM career orientation, but rather indirectly affects students' careers through self-efficacy and outcome expectations. Green and Sanderson (2018) stated that attending different programs such as summer field studies, high school outreach programs, and mentoring programs can increase STEM interest and prepare students for further study.

Expectancy

Lent et al. (1994, 2000) defined outcome expectations as the imaginary consequences of performing specific behaviors (e.g., what happens if I do this?). Many researchers have concluded that evidence regarding the effects of expectancies on STEM persistence is still lacking. In a longitudinal study of 710 high school participants, Aryee (2017) found that those with mid or higher levels of outcome expectations tended to persist and obtain college degrees in STEM. In another longitudinal study in which they employed the expectancy value model (Eccles et al., 2004) as a framework, the results showed that the expectations of low-income students were significantly influenced by pre-school contexts and directly affected post-secondary STEM persistence after three years of initial enrollment (Diemer & Li, 2012).

Given the importance of outcome expectancies, it is essential to address factors positively linked to them. Turner et al. (2019) examined the role of the outcome expectancies and found that mother support positively predicted students' outcome expectations and yielded career development in STEM. By contrast, Nugent et al. (2015) found that students' outcome expectations regarding STEM careers are directly influenced by interest and indirectly affected by educator, family, peer, and prior knowledge. However, limited research has addressed the influence of outcome expectations on students' postsecondary STEM persistence (Aryee, 2017). Also, the studies discussed above did not specifically explore the expectancies of high-achieving and non-high-achieving students.

Based on Social Cognitive Career Theory ([SCCT], Lent et al., 1994, 2000), Aryee (2017) concluded that outcome expectation is an essential part of the framework for predicting student persistence. Students with high level expectations are more likely to persist in STEM areas. Moreover, Heilbronner's (2011) study included factors that could shed light on the STEM persistence of students with talents; however, she did not include expectancy as a factor. Overall, it is important to note that research on how outcome expectancy affects students with gifts and talents and from diverse populations is limited.

Self-Efficacy

According to Bandura (1986), self-efficacy refers to an individual's beliefs in their capacity to perform the necessary behaviors to achieve specific accomplishments. The SCCT framework includes self-efficacy as an important factor in predicting several variables such as interests, goals, and persistence (Lent et al., 1994, 2000). Heilbronner (2011, 2013) found a significant correlation between self-efficacy and STEM degree retention. Dweck (2007) also pointed out that students' self-efficacy is one of the influential factors in STEM degree attainment. Furthermore, Nugent et

al. (2015) concluded that self-efficacy more strongly associated with STEM persistence when students' interests are reinforced by educators, peers, and family. According to Lewis et al. (2017), self-efficacy is positively associated with student achievement and motivation; nonetheless, the magnitude and quality of self-efficacy are different for male and female students. This results in gender-related differences favoring men in STEM persistence and attainments, which should be addressed by educators and researchers.

Gender

There is ongoing debate regarding the higher representation of men in STEM fields (Kim et al., 2018; Lewis et al., 2017; Turner et al., 2019, Wang & Degol, 2017) and the lower likelihood that women enter and continue in STEM areas to completion of degrees (Chen & Soldner, 2013; Heilbronner, 2013; Wang & Degol, 2017). Although Wang and Degol (2017) pointed out that the difference between male and female participation in STEM education has been declining, the U.S. Department of Commerce (2011) stated that the representation of females in the STEM workforce was only 25%, revealing that problems still exist concerning equal gender representation in STEM areas.

Researchers have investigated female students' low representation in STEM fields and found that gender gaps in STEM are not a result of cognitive abilities. However, Boston and Cimpian (2018) stated that low female representation in STEM is because of negative stereotypes against women. This includes intellectual abilities, confidence, sense of belonging, and interest. The authors pointed out that female students might feel less competent in STEM due to these negative stereotypes, and they might have low self-efficacy in STEM areas and interest.

Academic Achievement

Previous researchers have shown that domain specific abilities help students undertake and persist in their degrees (Kerr et al., 2012; Steenbergen-Hu & Olszewski-Kubilius, 2017). In many studies (e.g., Higdem et al., 2016) student achievements were measured through standardized test scores (e.g., SAT and AP exams), and in others GPA was used (e.g., Aryee, 2017; Camp et al., 2009).

With regard to gender, according to SAT Math and Verbal assessment results (College Board, 2011), male students have performed higher than female students since 1972. However, Heilbronner (2013) reported that male students' performance is more elevated on the SAT Math test but not on the Verbal test. Although ability is essential for success in STEM education, researchers have shown that achievement as measured by standardized tests is not effective in predicting female students' degree completion (Heilbronner, 2013).

Racial Background

Based on the literature, students of color (e.g., Black, Latinx, and Native American) face barriers (e.g., racism; low science and math grades) and are less likely to attend college (Ashford et al., 2016; Diemer & Li, 2012). Diemer and Li (2012), investigating the lower likelihood that students of color attend and attain degrees in postsecondary education, found that if age, academic achievement, and gender are controlled, peers and parents' educational expectancies support their college-level degree attainment. In a study of a U.S. national sample of Black male students with talents, Anderson (2016) found that 61% of eleventh-grade students who took pre-calculus in ninth- and eleventh-grades scored in the top 20%. However, only 18% of students who took only ninth-grade pre-calculus scored in the 20%. The result of the study further demonstrated that Black male students were more likely to persist in STEM if they attended field trips, engaged in

extracurricular activities, and took college level courses. Thus, involvement in STEM-related courses and activities promotes the pursuit and attainment of STEM degrees by high ability students from underrepresented populations.

Income

The low representation of students from economically disadvantaged backgrounds in postsecondary education is an ongoing issue as these students are more likely to drop out of school (Holmes et al., 2018). Turner et al. (2019) found that students with lower-income families perceived more barriers to careers in STEM and received less peer and parental and father support than higher SES peers, while the latter scored higher in outcome expectations but not in self-efficacy. Furthermore, the study results revealed that participating in STEM fields does not predict efficacy but does predict positive outcome expectations of students. More research is needed to determine the importance of income-related variables in STEM persistence for students from different economic backgrounds.

STEM Course Completion and Access

Preparation for and access to STEM courses is effective in increasing students' degree completion rates in STEM (Assouline et al., 2017). According to Anderson (2016), taking a higher-level mathematics course resulted in a higher graduation rate in any postsecondary field. Based on previous research, factors negatively associated with STEM degree completion are: fewer credit hours in STEM (Chen & Soldner, 2013), poor performance in STEM (Chen & Soldner, 2013). On the other hand, advanced math courses (Anderson, 2016), SAT math scores (Cardona et al., 2020; Chimka et al., 2007; French et al., 2005; Min et al., 2011; Nicholls et al., 2007; Thammasiri et al., 2014; Watkins & Mazur, 2013; Zhang et al., 2004), SAT verbal scores (Zhang et al., 2004), ACT

science scores (Chimka et al., 2007), high school GPA (French et al., 2005, Mendez et al., 2008; Nicholls et al., 2007; Watkins & Mazur, 2013; Zhang et al., 2004), freshman year GPA (Mendez et al., 2008), number of science courses taken (Mendez et al., 2008; Tyson et al., 2007; Wang, 2013), number of engineering courses taken (Mendez et al., 2008), number of math courses taken (Tyson et al., 2007; Wang, 2013), ACT and SAT scores (Nicholls et al., 2010), and 12th grade math achievement (Wang, 2013) positively affect STEM degree completion.

Advanced Placement (AP) and International Baccalaureate (IB) programs are two types of programs commonly offered to high school students looking for advanced curricula. The AP program has been offered since 1955 by the College Board (n.d.) to provide college-level introductory courses to students for those who desire more of a challenge. The AP program offers 38 different subjects, where students learn text examination, data interpretation, evidence evaluation, solid argument construction, and multiple perspectives. Students must obtain a 3 or higher grade to pass their AP test; thus, they can transfer credits to some colleges. Research has shown that students who take AP courses in high school are more likely to complete their STEM majors (Mattern et al., 2013).

However, students from underrepresented populations are less likely to have access to AP courses and are also underrepresented in the AP program (College Board, 2020). According to Smith et al.'s (2018) report, students from underrepresented populations in STEM, including first-generation students, students of color, and female students, had a 13% higher STEM college completion compared to those who did not take AP courses in high school. That demonstrates the need to support students from underrepresented populations to take AP courses to reduce potential dropouts in college.

The International Baccalaureate (IB) program was founded in 1968 by the International Baccalaureate Organization (2017) to promote more challenging courses and an internationally standardized curricula for students between 3 and 19 years old. IB programs are available in 158 countries, and the U.S offers 52% of the IB programs. IB programs increased by 37.9% between 2015 and 2019. Researchers (Pilchen et al., 2019) have shown that there is a positive relationship between students who are enrolled in an IB program and their persistence in postsecondary degrees in general.

Previous studies might help to understand student persistence; however, they do not study all STEM subjects as well as all grade levels. Therefore, there is a need to examine all possible STEM course offerings, which can yield a comprehensive picture of how these different courses and credits are associated with persistence.

CHAPTER 3 METHODS

In this study, I examined the demographic, cognitive, and non-cognitive factors affecting the early postsecondary STEM persistence of high-achieving students using nationally representative data from the High School Longitudinal Study of 2009 (HSLS: 2009). I used 2009 data as the base year, first follow-up data collected in 2012, update data collected in 2013, high school transcripts collected in 2013-2014, and second follow-up data collected in 2016. The HSLS: 2009 began with the data of a nationally representative ninth-grade cohort involving 944 schools in fall 2009. The first follow up involved data from the spring of 2012, when the students were in the eleventh grade. The second follow up involved their data three years after high school. Thus, students were followed through their high school and three of their postsecondary years. The HSLS: 2009 addresses how, when, and why students decide to pursue careers in STEM as well as in other areas.

In this study, I used the HSLS: 2009 data to investigate factors associated with early post-secondary STEM persistence in relation to the research questions using different approaches and compared how different machine learning methods address the research questions. I used feature selection techniques to select the best variables for the machine learning classification algorithms' performance in terms of predicting early post-secondary STEM persistence. Then, I used machine learning methods, including Random Forests and Artificial Neural Network, to find the selected variables' prediction accuracy. Finally, I used feature importance techniques to investigate the importance of each selected variable in the study.

Research Questions

This study addressed the following research questions:

1. What percentages of high-achieving and non-high-achieving students persist in postsecondary STEM majors?
2. What variables affect high-achieving and non-high-achieving students' persistence in postsecondary STEM majors?
3. Which variables most significantly influence the early postsecondary persistence of high-achieving and non-high-achieving students?
4. To what extent do high-achieving and non-high-achieving students' demographics (e.g., gender, ethnicity, socioeconomic status) affect their early postsecondary STEM persistence?
5. Which machine learning techniques can be used to identify variables influencing the early postsecondary persistence of high-achieving and non-high-achieving students in terms of classification models?

Data Sources

Data were drawn from the HSLS: 2009, which comprises nationally representative data collected by NCES. The HSLS: 2009 allows researchers to examine the transition of American students from high school to college or to the job market. The HSLS: 2009 is the fifth and only ongoing longitudinal study including a high school cohort by NCES. There are several new features included in the HSLS: 2009, which were not included in previous nationally representative longitudinal studies, such as the National Longitudinal Study of the High School Class of 1972, High School and Beyond, the National Education Longitudinal Study of 1988, and the Education

Longitudinal Study of 2002. For example, whereas previous studies started with eighth-, tenth- and twelfth-grade cohorts, in the HSLS:2009, the ninth-grade cohort was selected as the base because ninth grade is a critical juncture for to determining students transition to high school and academic paths (Ingels et al., 2011). Another difference is that student assessment, and student parent, teacher, and school administrator survey data are included in HLSL: 2009. Also, the HLSL: 2009 places particular emphasis on STEM, which allows researchers to address the relationships between STEM courses and achievement as well as persistence in STEM. The variables are selected from the publicly available data of the HSLS: 2009 to investigate the research questions of the study.

NCES Data Collection Procedures

The target population of the HSLS: 2009 comprised students in the United States, who were in ninth grade in the fall of 2009. The data were collected through a two-stage sampling process including a stratified random sampling of 1,889 schools of which a total of 944 schools agreed to participate in the study, and a random sample of 25,206 ninth-grade students (about 27 per school) (Ingels et al., 2011). The target population included students attending regular public schools, public charter schools, and private schools in the United States. Of the original sample of 25,206 students, 548 students who had language barriers or severe disabilities were excluded, resulting in 24,658 participants eligible to participate in the study. The student data comprised an assessment of algebraic reasoning and an online survey about educational experiences, expectancies, socioeconomic status, values regarding math and science as a subject, and vocations. Additionally, students' parents, principals, math and science teachers, and counselors completed surveys on the web or over the phone.

The first follow up took place 2.5 years after the base year data collection, in the spring of 2012, when the initial ninth-grade cohort of students were eleventh-graders. All students eligible to participate in the base year were included in the follow up data collection (Ingels et al., 2013). That is, the target population was not altered to have a representative cohort in the follow up. Therefore, 24,658 students were eligible to participate in the study in 2012.

The update data were collected between June and December 2013 to obtain information about students' transitioning from high school to college or the workforce. The purpose of the update was to examine students' high school completion, college applications and enrollment, and financial aid options status. Either a student or a parent completed a short 15-minute survey to provide information on students' high school completion, enrollment in courses for college credit, meetings with high school counselors, postsecondary enrollment, employment, financial aid, and careers (Ingels et al., 2015). As part of the update, the high school transcripts for the 2013-14 academic year of 23,415 students were collected from 938 of the initial 944 schools (six schools had closed after the baseline data collection). Students who had transferred to a new school were included in the follow up data.

The second follow up, which was administered from March 2016 through January 2017, included information on the target cohort approximately three years after high school graduation. This collection included new data such as high school completion; college enrollment history and plans, college majors, and occupations with an emphasis on STEM fields (Duprey et al., 2018). In this round of data collection, postsecondary transcripts were used instead of students' self-reports as a more reliable data source.

Participants

Identification

Students were identified for inclusion in this study as high-achieving and non-high-achieving based on their high school overall GPAs, which were included in the publicly available HSLS: 2009 dataset. In this study, all students (out of 23,503) with a 3.5 or higher GPA were identified as high-achieving, and non-high-achieving students were students with a 2.5 and lower GPA. I only selected students who were in STEM fields in college. Based on these criteria, 2,397 students identified as high-achieving in STEM and 1,034 students identified as non-high-achieving in STEM were included in the study to investigate early postsecondary STEM persistence predictors.

Of the students identified as high-achieving in this study, 17.73% were Asian, 4.01% were Hispanic, 6.76% were Multiracial, 0.58% were Indigenous, and 58.82% were White (Table 1). Of the students identified as non-high-achieving, 6.77% were Asian, 17.41% were Black, 19.83% were Hispanic, 9.38% were Multiracial, 1.84% were Indigenous, and 41.68% were White (See Table 1). It is important to note that 3.30% of high-achieving and 3.09% of non-high-achieving students' race information were missing in the dataset. More information on race proportions based on the HSLS: 2009 sample and National Population data is included in Table 1.

Table 1. Race Proportions Based on High-Achieving and Non-High-Achieving Students, HSLS: 2009 Sample, and National Population Census Data

	High-Achieving	Non-High-Achieving	HSLS: 2009	Nation
Asian	17.73%	6.77%	8.31%	5.90%
Black	4.01%	17.41%	10.42%	13.40%
Hispanic	8.80%	19.83%	16.16%	18.50%
Multiracial	6.76%	9.38%	8.26%	2.80%
Indigenous	0.58%	1.84%	1.17%	1.50%
White	58.82%	41.68%	51.41	60.10%
Missing	3.30%	3.09%	4.28%	

Note. The term Indigenous refers to students who are American Indian, Alaska Native, Native Hawaiian, and Pacific Islander. The national population is based on U.S Census Bureau's 2010 data.

Data Preparation

In the HSLS: 2009 dataset, missing values are encoded as certain numerical values. To handle these values, I assigned the "NaN" label to the missing values to prepare the data for data imputation. Also, I regrouped some ordinal categorical variables to have fewer categories related to the selected variable to increase generalization and learning capabilities of machine learning algorithms. Each feature in this study and its categories are explained in the variables section. Generally, datasets require addressing data quality problems such as removing duplicate values and handling missing values (Chu et al., 2016). However, in the HSLS: 2009 data, there was no problem regarding data cleaning since this process had already performed by the data provider, and the codebook was available for variables included in the dataset.

Data Imputation

Machine learning algorithms can work with missing values (Géron, 2019). However, missing values can negatively impact models' accuracy (Zahedi et al., 2020). In machine learning modeling, several methods are used for missing data imputation such as interpolation, mean,

median, constant value, forward fill (ffill), backward fill (bfill), and k-nearest neighbor (kNN). In this study, I used two different methods including ffill and kNN to fill the missing values since the selected variables contained categorical and numerical values. For simplicity, based on the final variables selected in the study, the number of missing values and their percentages are given for high-achieving and non-high-achieving students in the Tables 2 and 3, respectively.

Table 2. Missing Values Before Data Imputation for High-Achieving Students

Feature	Description	Count	Percentage
X1SEX	Gender	0	0
X1RACE	Race	79	3.32
X1SCHOOLBEL	School Belonging	203	8.53
X1SCHOOLENG	School Engagement	176	7.4
X1LOCALE	Locale	0	0
X2TXMPROF5	Math Proficiency	65	2.73
X2PAR1OCC_STEM1	Parent Occupation	110	4.62
X2BEHAVEIN	School Motivation	106	4.46
X2PROBLEM	School Problems	354	14.88
X2SCIEFF	Science Self-Efficacy	122	5.13
X2STU30OCC_STEM1	Expected Occupation at 30	98	4.12
X3TCREDCOMPSCI	Credits in Computer Science	0	0
X3TCREDAPIB	Credits in AP/IB combined	0	0
X3TGPASTEM	GPA in STEM	0	0
S1ACTIVITIES	Activity Attendance	0	0
X1SESQ5	Socioeconomic Status	150	6.31
X3TCREDSTEM	Credits in STEM	0	0
S4ANYDUALCRED	Dual Credits	131	5.51
Persistent		0	0

Table 3. Missing Values Before Data Imputation for Non-High-Achieving Students

Feature	Description	Count	Percentage
X1SEX	Gender	1	0.1
X1RACE	Race	32	3.09
X1SCHOOLBEL	School Belonging	117	11.32
X1SCHOOLENG	School Engagement	105	10.15
X1LOCALE	Locale	0	0
X2TXMPROF5	Math Proficiency	74	7.16
X2PAR1OCC_STEM1	Parent Occupation	87	8.41
X2BEHAVEIN	School Motivation	107	10.35
X2PROBLEM	School Problems	284	27.47
X2SCIEFF	Science Self-Efficacy	120	11.61
X2STU30OCC_STEM1	Expected Occupation at 30	86	8.32
X3TCREDCOMPSCI	Credits in Computer Science	0	0
X3TCREDAPIB	Credits in AP/IB combined	0	0
X3TGPASTEM	GPA in STEM	0	0
S1ACTIVITIES	Activity Attendance	0	0
X1SESQ5	Socioeconomic Status	77	7.45
X3TCREDSTEM	Credits in STEM	0	0
S4ANYDUALCRED	Dual Credits	87	8.41
Persistent		0	0

Categorical Missing Variables

Generally, machine learning researchers use several filling techniques for missing categorical values such as mode, ffill and bfill (Heydt, 2017). I used the ffill function for categorical missing values in the dataset, which uses the last valid observation forward to fill the missing value (Heydt, 2017). Thus, in this technique, the missing values are filled based on the last non-missing value in the previous row in the same column. This method helps prevent the overuse of one category and scatters values more randomly due to not assigning one specific value to all missing values (e.g., mode and median).

Numerical Missing Variables

In this study, I used the kNN algorithm for missing numerical values, which populates the values with respect to the selected k nearest neighbors. The algorithm uses a distance measure to fill the missing value similarly to surrounding values (Zhang, 2012). In this study, k is set as 15 to fill the missing numerical variables, indicating that missing values were filled relative to the nearest 15 neighbors. The kNN algorithm also helps avoid the overuse of one value in the same feature so that machine learning algorithms do not mislead learning. This method is widely used in machine learning due to its simplicity and high performance and because it is generally effective for numerical variables (Zhang, 2012).

Feature Scaling and Normalization

Machine learning algorithms underperform when the numerical input values have very different scales. Normalization (min-max scaling) is used to avoid this problem and does not change the shape of the distribution, only rescales values between 0 and 1 where the minimum value is equal to 0, and the max value is equal to 1. Using normalization is especially useful when the data are not normally distributed (Géron, 2019). Normalization places variables into a new scale (0-1). By doing this, machine learning algorithms do not place too much importance on features that have higher values. This scaling technique performs well with most of the machine learning algorithms including artificial neural networks as it does not assume any data distribution. Therefore, normalization is generally used with artificial neural networks (Géron, 2019).

Standardization

Standardization is another common scaling technique used in machine learning, and it is used to rescale values so that each feature has a mean of 0 and a standard deviation of 1, giving

the data a standard normal distribution (Géron, 2019). In this study, I only used normalization for numerical variables. I did not perform standardization because the numerical variables in this study had already been prepared and standardized by NCES based on statistical methods and were ready for use.

Feature Transformation and One-Hot-Encoding (Dummy Variables)

Categorical variables need to be encoded with numerical values to be meaningful in machine learning modeling. However, when we assign categorical variables with numerical values, machine learning algorithms may assume an ordinal relationship; such a relationship can damage the model. Furthermore, in machine learning algorithms, two near values are considered more similar than two distant values (Géron, 2019). The most common way to prevent these problems is to employ the one-hot-encoding or one-out of-N encoding, which is also called dummy variables (Müller & Guido, 2016). This method replaces integer categorical variables with new attributes into one-hot vectors, which contain values of 0 and 1 (Géron, 2019; Müller & Guido, 2016). The number of these new attributes are based on the categories of the categorical variable. In this study, I used one-hot-encoding for all categorical variables with three or more categories to obtain dummy variables for each category.

Variables

In machine learning, “feature” is the term used to describe a variable. In this study, the variables of interest, including demographics, cognitive factors, and non-cognitive factors, are defined and discussed in the following paragraphs.

Sex

In this study, I regrouped the "X1SEX" variable, and it included categories of "0" and "1" for male and female students, respectively.

Race

In the HSLS: 2009 dataset, race/ethnicity is a composite variable encompassing six different dichotomous race/ethnicity composites. In this study, I regrouped the categories of the "X1RACE" variable based on the publicly available dataset and used the following categories in the analysis: Asian, Black/African American, Hispanic/Latino, Indigenous (including American Indian, Alaska Native, Native Hawaiian, and Pacific Islander), Multiracial, and White. In the HSLS: 2009, race/ethnicity composites were obtained through a student questionnaire, and if race information was not available on the student questionnaire, a school-provided sampling roster or data from a parent questionnaire was used to generate the race ethnicity composites.

Parental Education

In the HSLS: 2009, several categorical variables are included for parents' education. In this study, I included all parental education-related variables and aimed to select the final variables after using the feature selection techniques to identify those with greatest influence on persistence. Therefore, I selected variables from each parent's highest level of education, both parents' highest level of education, female guardian's highest level of education, and male guardian's highest level of education in this study. Based on the feature selection analyses, there was a high correlation between parental education variables. Furthermore, a socioeconomic status variable was created using parental education so that I decided to avoid using parental education related variables in my final analysis as it did not bring new information into the analyses.

Socioeconomic Status (SES)

For the socioeconomic status, I used the quintile variable "X1SESQ5", which consists of 5 quintiles from 1 to 5, from the lowest to the highest quartile, respectively. The quintile variable for the socioeconomic status is created by NCES for inclusion in the dataset using both the parents'/ guardians' education, occupation, and family income.

Math Scores

In this study, I used several math related variables to select the best predictive math scores associated with STEM persistence in college. I added the following variables: math quantile score, base year math proficiency probability scores from level 1 to 5, and follow up math probability scores from level 1 to 7. Each score is created by NCES and explained in more detail in the following paragraphs.

Math Quantile Score

This is a norm-referenced achievement variable included in the HSLS: 2009 dataset. This score is derived from the estimated (weighted) population achievement distributions based on math scores, where “1” indicated the lowest and “5” indicated the highest quantile. These quantiles are generated through the cut points at every 20^{th} percentile.

Math Proficiency Probability Scores (Base Year)

In this study, five different levels of probability scores were used. These scores are criterion referenced and clustered using IRT-estimated item parameters created by NCES. The higher a score the lower the level of proficiency, and a student who is at a particular proficiency is expected to pass that given level. In the HSLS: 2009, levels 1 through 5 represent proficiency in

the following math topics: algebraic expressions, multiplicative and proportional thinking, algebraic equivalents, systems of equations, and linear functions, respectively.

Math Proficiency Probability Scores (First Follow up)

Similar to the math proficiency probability scores of the base year, the first follow up math proficiency scores are criterion referenced and based on clusters of items that mark seven levels on the mathematics scale. In HSLS: 2009, the levels are hierarchical so that higher level mastery subsumes proficiency at the lower levels. These proficiency scores were computed by NCES using IRT-estimated item parameters, and each of the proficiency scores consists of a continuous scale. Each proficiency probability represents the probability that a student would pass a given proficiency level. In the first math proficiency follow-up scores, levels 1 to 7 correspond to the following math topics: algebraic expressions, multiplicative and proportional thinking, algebraic equivalents, systems of equations, linear functions, quadratic functions, and log and exponential functions, respectively.

Identity

Math Identity

I included both the base year and the first follow up math identity variables in the study. In the HSLS: 2009, math identity is a continuous variable that was generated by NCES using principal components factor analysis and standardized to a mean of 0 and standard deviation of 1. Students who agree with the statements ″you see yourself as a math person" and/or "others see me as a math person″ have higher values for the math identity.

Science Identity

Similar to math identity, this variable is a continuous variable for science identity and constructed from extent of agreement with the following statements: ″you see yourself as a science person" and/or "others see me as a science person." This variable was created by NCES using factor analysis and standardized scores. In the HSLS: 2009, base year and first follow up science identity variables were generated, and these variables were included in this study.

Interest

Math Interest

In the HSLS: 2009, this variable represents students' math interest on a scale in which higher values represent greater interest in math courses. This variable was created by NCES using principal components factor analysis and standardized to a mean of 0 and standard deviation of 1. The scale is generated based on six input variables, three of which are assessed with four-point Likert scale items: *"*you are enjoying/enjoyed this class very much or you enjoy math classes very much," "you think/thought this class is/was a waste of your time or you think math classes are a waste of your time," and "you think/thought this class is/was boring or you think math classes are boring." The remaining three input variables were related to the items "favorite school subject," "least favorite school subject," and "taking fall 2009 math because he/she really enjoys math." Students who provided a full set of responses were assigned a scale value for math interest.

Science Interest

This variable is also a continuous variable, which is generated using the same procedures as those used with math interest. A higher value represents a greater interest in science courses. The variable was created by NCES and constructed using factor analysis and standardized scores.

This scale is based on the following statements: *"*enjoying fall 2009 science course very much," "thinks fall 2009 science course is a waste of time," "thinks fall 2009 science course is boring," "favorite school subject," "least favorite school subject," and "taking fall 2009 science because he/she really enjoys science."

Self-Efficacy

Math Self-Efficacy

Math self-efficacy is a continuous variable in which a higher value represents higher math self-efficacy. In the HSLS: 2009, math self-efficacy was created using principal components factor analysis and standardized scores from the base year data. This scale was generated using four input variables, each measured using a four-point Likert scale: "you are confident that you can do an excellent job on math assignments," "you are certain that you can understand the most difficult material presented in math textbooks," "you are certain that you can master math skills," and "you are confident that you can do an excellent job on math assignments." Students who provided all responses were assigned a scale value.

Additionally, the first follow up math self-efficacy scale was also included in this study, which is generated using the same calculation methods and inputs that were used in the first follow up data for the scale score.

Science Self-Efficacy

In the HSLS: 2009, this variable is a scale of students' science self-efficacy in which a higher value represents higher science self-efficacy. The variable was obtained from the base year data and generated through principal components factor analysis and standardized scores. Four input variables, each presented as a four-point Likert scale item, were used to compute this scale:

"you are confident that you can do an excellent job on science tests," "you are certain that you can understand the most difficult material presented in science textbooks," "you are certain that you can master science skills," and "you are confident that you can do an excellent job on science assignments." Students who provided a full set of responses were assigned a scale value.

In the HSLS: 2009, the first follow up science self-efficacy scale, a similar variable to the science self-efficacy score from the base year data, was constructed using the same calculation methods and inputs for the scale score.

Sense of Belonging

This variable represents students' perceptions of school belonging, in which higher values represent a greater sense of school belonging. The sense of belonging variable was generated through principal components factor analysis and standardized scores. Five inputs, each presented as an item with a four-point Likert scale, were selected to generate the sense of belonging scale: *"*you feel safe at this school," "You feel proud being part of this school," "there are always teachers or other adults in your school that you can talk to if you have a problem," "school is often a waste of time," and "getting good grades in school is important to you." Students who provided a full set of responses were assigned a scale value for sense of belonging.

Engagement

In the HSLS: 2009, the engagement variable is represented on a scale of the students' school engagement in which a higher value represents greater school engagement. The variable was calculated through principal factor components and standardized scores. In this study, to generate the engagement variable, four input variables were measured with the following questions: "how often do you go to class without your homework done?" "how often do you go to

class without pencil or paper?" "how often do you go to class without books?" and "how often do you go to class late? Only students who responded to all these questions were assigned a scale value in the HSLS: 2009.

Motivation

In the HSLS: 2009, there is a numerical value that represents a scale of students' school motivation. The variable was created using principal components factor analysis and standardized to a mean of 0 and standard deviation of 1 by NCES, and higher values represent higher school motivation.

Student Expectation

This is a categorical variable indicating the highest level of education each ninth grader expects to achieve. This variable is drawn from the student questionnaire, and the same variable was collected in the first follow up and was also included in this study. In this study, I regrouped categories of student expectation as follows: "do not know," "high school," "college degree," and "graduate degree."

Parent Expectation

Another categorical variable included was "how far in school a ninth grader's parents thinks s/he will go." This variable is derived from the base year parent questionnaire, or, if missing from the base year parent questionnaire, it is statistically imputed in the HSLS: 2009. Additionally, the same variable was collected in the first follow up and was also included in this study.

School Problems

In the HSLS:2009, there is a standardized value that represents a scale of problems such as lack of resources and materials at each high school. The variable was created by NCES using principal components factor analysis and standardized to a mean of 0 and standard deviation of 1 by NCES and higher values represent more positive assessments of the school's problems.

School Locale and Region

School Locale

School locale is a categorical variable based on four categories: city, suburb, town, and rural represented by values 1 to 4, respectively. Values for the base year, the first follow up, and the update were included in this study. The update data demonstrate the locale of students' current or last attended school or "-8" for unit non-response and "-9" for missing values.

School Region

School region is a categorical variable indicating the geographic location of a student's school in the United States, Northeast, Midwest, South, and West, represented by values 1 to 4, respectively. Values for the base year, the first follow up, and the update were included in this study. The update data demonstrates the geographic location of the sample based on students' current or last attended school also includes "-8" for unit non-response and "-9" for missing values.

Math and Science Effort

This math effort variable is a scale of the students' responses to math effort items; higher values represent more positive assessments. This variable was generated using factor analysis and standardized scores. To generate this scale, four input variables were used, each consisting of a

five-point Likert scale item: “you pay/paid attention to the teacher,” “you turn/turned in your assignments and projects on time,” “when an assignment is/was very difficult, you stop/stopped trying,” and “you do/did as little work as possible; you just want/wanted to get by.”

In the HSLS: 2009, the science effort scale was created by following the same procedures as for the math effort scale, the only difference being science was referenced instead of math.

Activities

In this study, I created a variable based on students’ responses, representing the participation of any of the following student activities including math club, math competition, math camp, math study groups, science club, science competition, science camp, and science study groups in 2008-2009 school year. The variable consisted of two categories, representing “0” for no and “1” for yes.

Courses and Credits

In the HSLS: 2009, several variables are included for the credits earned in STEM courses and the highest-level STEM courses taken. I selected all variables regarding STEM courses and credits in this study and choose the final variables after using feature selection techniques.

Dual Credit

This is a dichotomous categorical variable representing if the student ever earned dual enrollment credits from college or trade school. The variable consisted of two categories, representing “0” for no and “1” for yes.

GPA in STEM

This variable is a numerical value (max GPA is 4, min GPA is 0), and it represents total GPA earned in STEM courses in high school. The variable was included in the high school transcripts collected in 2013-2014.

Persistence in STEM Fields

In this study, persistence is defined in terms of students who continue their education in STEM fields within 3 years of initial college enrollment (including all types of postsecondary education institutions). In other words, if they start a STEM degree and continue in a STEM degree, they are considered persistent in STEM. The dependent variable of the study is persistence in STEM fields. I used different variables to create a binary outcome (target) variable (feature). This procedure is called feature engineering, which allows for extracting different features from the dataset.

To create the target variable, I used a variable to indicate whether or not students initially considered pursuing a STEM major drawing on information from the second follow up and using the values of "0" for no and "1" for yes. Then, I used a categorical variable that indicates students' first or second major fields of study in STEM areas. In the HSLS: 2009, this variable indicates whether a student is enrolled in a STEM field or not. The variable includes two values, "0" for not having a STEM major and "1" for having one or more STEM majors. Another variable used in this study is also categorical, indicating students' NSF STEM majors, indicated as "0" for not having an NSF approved STEM major and "1" for having an NSF approved STEM major. In this study, if a student in a STEM or NSF STEM, that is considered as a STEM major. Ultimately, students who initially considered pursuing STEM majors and still pursue STEM majors and are considered as persistent. The remaining students who did not continue STEM majors are

considered as non-persistent in this study. The persistency and non-persistency categories constituted the dependent/target/outcome variable in this study.

Feature Selection

The outcome variable of the study was a dichotomous variable that demonstrated whether students persisted in STEM majors within three years of college enrollment. In this study, I used feature selection techniques to select the best variables for early postsecondary STEM student persistence and reached the models used in the study. I aimed to explain the model with the best predictive variables to explain students' persistence in STEM adequately.

Numerical Variables

In terms of numerical variable selection, Pearson Correlation and Variance Inflation Factor (VIF) values were used in the study. I grouped all numerical variables in the study and examined the Pearson Correlation. Pearson correlation uses coefficients ranging from -1 to 1, where "-1" and "1" represent strong negative and positive correlations, respectively. I investigated variables with moderate and strong correlations in more detail and selected final variables according to the literature and correlation results.

I examined the variance inflation factor of the remaining variables after Pearson Correlation. I observed the VIF of all the remaining variables, and they were less than 2.5 and lower, indicating there was no multicollinearity among variables (Midi et al., 2010).

Categorical Variables

Before using categorical feature selection, I encoded categorical variables consisting of 3 or more categories to obtain dummy variables for each category. Next, I examined Spearman Correlation and Chi-Square values.

In this study, I also used Spearman Correlation that uses coefficients ranging from -1 to 1, where "-1" and "1" represent strong negative and positive correlations, respectively. In this study, I examined variables with moderate and strong correlations in more detail and obtained the final variables according to the literature and correlation results. In this study, no variables were reduced using the chi-square; this method is used for control purposes.

Feature Importance

Feature importance techniques use unique machine learning parameters from the relevant machine learning libraries to select the most important features for the model (Saeys et al., 2008). The most important features are good predictors of the independent variable (target feature). After feature selection, I employed feature importance techniques for identifying good predictors of persistence in STEM. Each feature importance technique uses different machine learning algorithms and parameters to provide a score for each independent variable. Higher scores represent variables that are more salient to the dependent variable. For instance, if the feature importance score is high for a specific feature in this study, the feature affects students' persistence more than other features. These techniques reduce the complexity of the model (number of features or independent variables) so that interpretations become easier. I used the random forest feature importance technique to examine each variable's importance. The higher scores represent the higher importance for the target variable. In this technique, a sum of all importance scores is equal to 1.

Machine Learning Methods

Random Forest

Random forest is an ensemble learning method and consists of decision tree collections (Müller & Guido, 2016). A random forest incorporates the random selection of variables (Ho, 1995; Amit & Geman, 1997). In other words, in a random forest, a subset of random variables is selected in the learning process. According to Hastie et al. (2009), for a classification problem with p variables, the size of random variables is typically set at $\sqrt{p}$; for a regression problem, the size p is set at $\frac{p}{3}$. In addition to the number of random variables, a random forest also includes other tuning parameters, including the terminal node size, number of trees, and bootstrap sample sizes.

The random forest method has some advantages, such as providing high quality models, reducing overfitting issues, and selecting the most important dataset variables. In this study, I used several parameters in random forest. The "max samples" parameter was used with bootstrap that was set as "True" to control the decision tree collections. Furthermore, I used "n estimators" as 1000, indicating the number of trees in the forest. The quality of a split was measured using Gini impurity. All the remaining parameters were used as default. I also set "random state" as 46 in the study to produce the same results while using the random forest method. In this study, I randomly selected 80% of data for training and 20% of data for testing.

Artificial Neural Networks (ANN)

The artificial neural network is a non-parametric machine learning technique in which many simple units, called neurons, are interconnected by weighted links (Mason et al., 2018). The human brain inspired artificial neural networks, and artificial neural network neurons work

similarly to neurons in the human brain (Rosebrock, 2019). The artificial neural network is a labeled structure and consists of an input layer, weights, hidden layers, and an output layer (Rosebrock, 2019). Each layer has nodes, and connections through layers are obtained through signals. A final function calculates the output label. According to Mason et al. (2018), artificial neural networks outperform standard classical methods because they can analyze incomplete and noisy data, and they provide clear solutions.

In this study, the artificial neural networks used included 1 input layer, 3 hidden layers, and 1 output layer. As an activation function in each hidden layer, I used Rectified Linear Units (ReLU) to introduce nonlinearity into the artificial neural network to increase learning capabilities with nonlinear learning functions. Using the output of the last layer, sigmoid classifies student persistency into two different classes based on the highest probability.

The data trained with using neural network algorithm with a batch size of 10 examples, the learning rate of 1e−2 (scientific notation of .01). The artificial neural network Keras model compiled with the "binary_crossentropy" loss function, "adam" optimizer, and "accuracy" metrics using the Keras library (Chollet et al., 2015). Depending on the amount of data (e.g., number of high-achieving students' data) and selected features in the training and test sets, the max epoch number in experiments was set to 100 and 125.

Evaluation Measures

To evaluate the performance of machine learning methods, I used accuracy, sensitivity, and specificity metrics in the study. These evaluation metrics are calculated using a confusion matrix (also known as error matrix) representing actual vs. predicted results in a table format similar to the decision table for hypothesis testing. Because the study outcome variable had only

two categories (dichotomous outcome), there were four populated cells in the confusion matrices in Table 4.

Table 4. Confusion Matrix for a Dichotomous Outcome Variable

		Predicted Results	
		Predicted positive	Predicted negative
Actual Results	Actual positive	True Positive (TP)	False Positive (FP)
	Actual negative	False Negative (FN)	True Negative (TN)

Accuracy

Accuracy is the amount of correctly predicted data out of all the data. In this study, accuracy shows the percentage of correctly predicted data by the machine learning algorithms, which is the amount of correctly predicted data divided by all the data in the dataset. The amount of correctly predicted data can be defined as follows: the total number of class 0 classified as class 0 (True Negative) and the total number of class 1 classified as class 1 (True Positive).

The terms True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) refer to the result of a test and the correctness of the classification. For example, in this study, for student persistence, TP means correctly predicted as persistent, FP means incorrectly predicted as persistent, TN means correctly predicted as not persistent, and FN is incorrectly diagnosed as persistent.

Accuracy is calculated as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP + TN + FP+ FN}$$

Sensitivity

Sensitivity, also known as recall, is another common evaluation measure used in machine learning studies. Sensitivity is the proportion of actual positive cases that are predicted as positive (True Positive) and provides information on how well the test detects True Positives (TP).

Sensitivity is calculated as follows:

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

Specificity

Specificity is defined as the proportion of the actual negative class (non-persistent), which is predicted as the negative class (True Negative). Thus, specificity provides information on how well the test detects True Negatives (TN).

Specificity is calculated as follows:

$$\text{Specificity} = \frac{TN}{TN+FP}$$

Receiver Operating Characteristics (ROC) Curve

A *ROC curve* is a plot showing the tradeoff between sensitivity and specificity. A ROC curve is a graphically illustrated plot that shows the ability of a binary classifier system as its discrimination threshold is varied. Normally, the threshold value is defined as 0.5 as a default which means if the classification result is greater than and equal to 0.5, it is predicted as positive class (persistent). With ROC curves, we can see the machine learning algorithm performance when the threshold increases or decreases. When the true positive rate increases and the false positive rate decreases, the area under the ROC curve enlarges, indicating that the test is more accurate. If

the ROC area becomes a 45-degree diagonal, the test is less accurate. In sum, the area under the ROC curve indicates the test's accuracy.

Handling Imbalanced Classification

Synthetic Minority Over-Sampling Technique (SMOTE)

Imbalanced learning is a common problem in machine learning, requiring special attention for minority class (e.g., students who are not persistent in STEM fields in this study). There are several approaches to handle imbalanced learning to avoid misclassification of minority class such as under-sampling and over-sampling (Longadge et al., 2013). In this study, I performed Synthetic Minority Over-Sampling Technique (SMOTE), which generates additional data points based on the sampling data of minority class using K-nearest neighbors (Elreedy & Atiya, 2019). SMOTE is a common technique in machine learning to prevent the imbalanced learning problem (Elreedy & Atiya, 2019).

CHAPTER 4 RESULTS

In this section, I present the results of the study, in which I investigated high-achieving and non-high-achieving students' persistence in postsecondary STEM fields using random forest and artificial neural network methods. I also discuss the importance of each selected factor affecting students' persistence using random forest feature importance technique.

Research Question 1: What percentages of high-achieving and non-high-achieving students persist in postsecondary STEM majors?

To answer the first research question, I used publicly available student data from HSLS: 2009. The full dataset includes 23,503 students. To perform the analysis, I created "Major" and "Persistent" variables by using feature engineering techniques. For high-achieving and non-high-achieving students, I selected STEM students with GPAs of 3.5 and higher and 2.5 and lower, respectively. A total of 2,397 students were identified as high-achieving and 1,034 students as non-high-achieving.

In this study, 29.62% of high-achieving students were non-persistent while 70.38% were persistent in STEM fields. On the other hand, 28.92% of non-high-achieving students were non-persistent, while 71.08% were persistent in STEM majors. These results indicate that the percentages of high-achieving and non-high-achieving of students who were persistent in STEM majors were similar. However, the number of high-achieving students who pursued STEM were two times more than non-high-achieving students.

Research Question 2: What variables affect high-achieving and non-high-achieving students' persistence in postsecondary STEM majors?

To select variables that affect high-achieving and non-high-achieving students' persistence in postsecondary STEM majors, I used a dataset that included approximately 8,500 student variables. Based on the literature, I included all the variables that were relevant to student persistence in STEM in my initial analyses. In this study, a dichotomous outcome variable (target feature), created using feature engineering techniques, indicated whether students persisted in their initially designated STEM major. I employed feature selection techniques to select the most important features for early post-secondary student persistence in STEM and reached optimal performance of the machine learning models used in this study. I aimed to explain the model with the variable that best predicted student persistence in STEM.

I used two different approaches to select final numerical variables, Pearson Correlation and the Variance Inflation Factor (VIF). Variables with moderate to high Pearson correlations were examined in more detail, and the final variables were selected based on the domain knowledge and correlation results. Next, I checked the remaining variables' VIFs. I observed that all the remaining variables' VIFs were 2.5 or lower, which indicated no multicollinearity among variables.

Additionally, I encoded categorical variables that consisted of three or more categories and obtained dummy variables for each category; this method is also known as one-hot-encoding. Then, in order to select the final categorical variables, I used Spearman Correlation and Chi-Square methods. Variables with moderate to high Spearman correlations were examined in more detail. The final variables were selected based on the correlation results. For example, the total GPA was used instead of each grade level GPA because the total GPA provided more information and correlated with each grade level GPA. I did not remove any variables based on chi-square because the chi-square value for each feature provided no new information. I used this method for checking

purposes as I had observed that all the variables were correctly selected based on the Pearson Correlation Analysis.

After my initial analyses, I selected the following variables for the study: school belonging, school engagement, math proficiency level, parent occupation, school motivation, school problems, science self-efficacy, expected occupation at the age of 30, credits earned in computer sciences, GPA for STEM courses, student activities, credits earned in STEM, credits earned in AP/IB combined, dual credits, race, socioeconomic status, urbanicity. Information on all variables used in the analyses is included in Table 5.

Table 5. Full List of Final Variables

Feature	Description	Type	Values
X1SEX	Gender	Categorical (Binary)	0: Female, 1: Male
X1SCHOOLBEL	School Belonging	Numerical	Normalized float values ranging from 0 to 1
X1SCHOOLENG	School Engagement	Numerical	Normalized float values ranging from 0 to 1
X2TXMPROF5	Math Proficiency	Numerical	Normalized float values ranging from 0 to 1
X2PAR1OCC_STEM1	Parent Occupation	Categorical (Binary)	0: Non-STEM, 1: STEM
X2BEHAVEIN	School Motivation	Numerical	Normalized float values ranging from 0 to 1
X2PROBLEM	School Problems	Numerical	Normalized float values ranging from 0 to 1
X2SCIEFF	Science Self-Efficacy	Numerical	Normalized float values ranging from 0 to 1
X2STU30OCC_STEM1	Expected Occupation at 30	Categorical (Binary)	0: Non-STEM, 1: STEM
X3TCREDCOMPSCI	Credits in Computer Science	Numerical	Normalized float values ranging from 0 to 1
X3TCREDAPIB	Credits in AP/IB combined	Numerical	Normalized float values ranging from 0 to 1
X3TGPASTEM	GPA in STEM	Numerical	Normalized float values ranging from 0 to 1
S1ACTIVITIES	Activity Attendance	Categorical (Binary)	0: No, 1: Yes
X3TCREDSTEM	Credits in STEM	Numerical	Normalized float values ranging from 0 to 1
S4ANYDUALCRED	Dual Credits	Categorical (Binary)	0: No, 1: Yes
X1RACE_2.0	Asian	Categorical (Binary)	Asian 0: No; 1: Yes
X1RACE_3.0	Black	Categorical (Binary)	0: No, 1: Yes
X1RACE_5.0	Hispanic	Categorical (Binary)	0: No, 1: Yes
X1RACE_6.0	Multiracial	Categorical (Binary)	0: No, 1: Yes
X1RACE_7.0	Indigenous	Categorical (Binary)	0: No, 1: Yes
X1RACE_8.0	White	Categorical (Binary)	0: No, 1: Yes
X1LOCALE_1.0	City	Categorical (Binary)	0: No, 1: Yes
X1LOCALE_2.0	Suburb	Categorical (Binary)	0: No, 1: Yes
X1LOCALE_3.0	Town	Categorical (Binary)	0: No, 1: Yes
X1LOCALE_4.0	Rural	Categorical (Binary)	0: No; 1: Yes
X1SESQ5_1.0	First quintile (lowest)	Categorical (Binary)	0: No; 1: Yes
X1SESQ5_2.0	Second quintile	Categorical (Binary)	0: No; 1: Yes
X1SESQ5_3.0	Third quintile	Categorical (Binary)	0: No, 1: Yes
X1SESQ5_4.0	Fourth quintile	Categorical (Binary)	0: No, 1: Yes
X1SESQ5_5.0	Fifth quintile (highest)	Categorical (Binary)	0: No, 1: Yes
Persistent		Categorical (Binary)	0: No, 1: Yes

Note. The term Indigenous refers to students who are American Indian, Alaska Native, Native Hawaiian, and Pacific Islander.

Research Question 3: Which variables most significantly influence the early postsecondary persistence of high-achieving and non-high-achieving students?

To examine variables that most significantly influenced high-achieving and non-high-achieving students' early postsecondary persistence, I used different feature importance techniques. Based on the machine learning algorithm performances, I obtained best results with random forest and artificial neural network methods using a Synthetic Minority Oversampling Technique (SMOTE) dataset. As the random forest importance technique is one of the commonly used methods in educational research for non-linearly distributed data, I used it. The results for high-achieving and non-high-achieving students are listed in Table 6. The random forest importance algorithm uses random forest classifiers and assigns a relative importance score for each variable. According to this technique, the higher the percentage of the variable, the greater its importance.

Table 6. Feature Importance Scores of High-Achieving (GPA >= 3.5) and Non-High-Achieving STEM Students (GPA <= 2.5)

	High-Achieving Students		Non-High-Achieving Students	
Feature	Feature Importance Score	Percentage	Feature Importance Score	Percentage
Gender	0.017706	1.77	0.025039	2.5
School Belonging	0.078048	7.8	0.075099	7.51
School Engagement	0.085393	8.54	0.077479	7.75
Math Proficiency	0.09165	9.17	0.089296	8.93
Parent Occupation	0.012636	1.26	0.019495	1.95
School Motivation	0.083039	8.3	0.086969	8.7
School Problems	0.082953	8.3	0.076305	7.63
Science Self-Efficacy	0.07525	7.53	0.073064	7.31
Expected Occupation at 30	0.017334	1.73	0.028278	2.83
Credits in Computer Science	0.066012	6.6	0.067485	6.75
Credits in AP/IB combined	0.068642	6.86	0.033737	3.37
GPA in STEM	0.078694	7.87	0.075965	7.6
Activity Attendance	0.012793	1.28	0.012708	1.27
Credits in STEM	0.074368	7.44	0.072674	7.27
Dual Credits	0.014915	1.49	0.014777	1.48
Asian	0.010782	1.08	0.007697	0.77
Black	0.004655	0.47	0.012142	1.21
Hispanic	0.008127	0.81	0.014319	1.43
Multiracial	0.006979	0.7	0.008672	0.87
Indigenous	0.001021	0.1	0.002565	0.26
White	0.013728	1.37	0.013518	1.35
City	0.012215	1.22	0.01403	1.4
Suburb	0.013753	1.38	0.014306	1.43
Town	0.009647	0.96	0.008878	0.89
Rural	0.011669	1.17	0.011762	1.18
First quintile (lowest)	0.005516	0.55	0.010347	1.03
Second quintile	0.007967	0.8	0.013763	1.38
Third quintile	0.009572	0.96	0.016529	1.65
Fourth quintile	0.011739	1.17	0.012531	1.25
Fifth quintile (highest)	0.013198	1.32	0.010569	1.06

Note. Feature importance scores are based on SMOTE datasets.

Based on the feature importance results for high-achieving students, the 10 most important variables for the study were math proficiency level 5 (9.17%), school belonging (7.80%), school engagement (8.54%), school motivation (8.30%), school problems (8.30%), science self-efficacy (7.53%), credits earned in computer sciences (7.44%), GPA for STEM courses (7.87%), credits earned in STEM courses (7.44%), and credits earned in AP/IB combined (6.86%). Based on these results, math proficiency was the most important variable in the study, showing the greater importance of math proficiency in predicting persistence of high-achieving students in STEM fields.

Feature importance results for non-high-achieving students (Table 6) revealed that the 10 most important variables of the study were math proficiency level 5 (8.93%), school motivation (8.70%), school engagement (7.75%), school problems (7.63%), GPA for STEM courses (7.60%), school belonging (7.51%), credits earned in STEM courses (7.27%), science self-efficacy (7.31%), credits earned in computer sciences (6.75%), and credits earned in AP/IB combined (3.37%).

According to these results, the most important variables were similar for high-achieving and non-achieving students. However, the importance scores of these variables were different. For example, the scores for credits earned in AP/IB were 3.37% for non-high-achieving students and 6.86% for high-achieving students, meaning that the variable was almost twice as important for high-achieving students. Furthermore, when I compared academic factors in general (see Table 6), I found that academic factors had slightly higher importance for high-achieving students' persistence in STEM.

Research Question 4: To what extent do high-achieving and non-high-achieving students' demographics (e.g., gender, ethnicity, socioeconomic status) affect their early postsecondary STEM persistence?

Table 7 presents the demographic information (gender, race, socioeconomic status, and locale) of high-achieving STEM students shows that 68.04% of female and 73.73% of male high-achieving students were persistent in STEM fields; however, the number of female students (927) in this group was greater than that of male students (727). Results regarding race showed that White (71.48%), Hispanic (70.59%), and Asian (70.53%) students had the highest percentages of persistent high-achieving students in STEM areas. The proportion of persistent high-achieving Multiracial students was 65.09% and of Black students was 64.65%. Indigenous students had the smallest percentage of persistent high-achieving students, 53.33%. Furthermore, in this study, comparison of the percentages of persistent students from different socioeconomic backgrounds showed that the highest percentage of persistent students who were high-achieving in STEM fields were from the third economic quintile (73.21%), followed by the fourth (72.50%), fifth (69.41%), and second (68.37%) quintiles (see Table 7). The smallest percentage of persistent students was from the first (67.16%) quintile, indicating that students from lower socioeconomic backgrounds were the least persistent in STEM. Regarding students' residential locale, the highest percentage of persistent high-achieving students (72.85%, n=884) resided in suburbs. In contrast, the lowest percentage of persistent high-achieving students (64.26%, n=291) resided in towns. There is not a notable difference in student persistence based on students' residential locale.

As shown in Table 7's summary of demographic information for non-high-achieving students, 69.94% of female and 72.28% of male non-high-achieving students were persistent in STEM. Among racial groups, Hispanic students had the highest percentage of persistence in STEM fields (76.53%) followed by White students (68.62%), and Indigenous students had the lowest percentage (57.89%). With regard to socioeconomic status, the highest percentage of persistent

non-high-achieving students were from the third quintile (75.41%) and the lowest percentage (64.84%) from the fifth quintile (see Table 7). Regarding urbanicity, the highest percentage of persistent non-high-achieving students was from rural areas (73.33%, *n*=165) and the second highest from suburban areas (72.52%, *n*=285), followed by city (68.89%, *n*=217) and town (67.33%, *n*=68).

Table 7. Demographics of High-Achieving (GPA >= 3.5) and Non-High-Achieving (GPA <= 2.5) STEM Students

		High-Achieving Students			Non-High-Achieving Students		
Variable	Categories	Total Number	Non-Persistent	Persistent	Total Number	Non-Persistent	Persistent
Gender	Female	1411	451 (31.96 %)	960 (68.04%)	529	159 (30.06%)	370 (69.94%)
	Male	986	259 (26.27%)	727 (73.73%)	505	140 (27.72%)	365 (72.28%)
Race	Asian	431	127 (29.47%)	304 (70.53%)	72	21 (29.17%)	51 (70.83%)
	Black	99	35 (35.35%)	64 (64.65%)	189	52 (27.51%)	137 (72.49%)
	Hispanic	221	65 (29.41 %)	156 (70.59%)	213	50 (23.47%)	163 (76.53%)
	Multiracial	169	59 (34.91 %)	110 (65.09%)	98	29 (29.59%)	69 (70.41%)
	Indigenous	15	7 (46.67%)	8 (53.33%)	19	8 (42.11%)	11 (57.89%)
	White	1462	417 (28.52%)	1045 (71.48%)	443	139 (31.38%)	304 (68.62%)
Socioeconomic Status	First Quintile	134	44 (32.84%)	90 (67.16%)	196	49 (25%)	147 (75%)
	Second Quintile	215	68 (31.63%)	147 (68.37%)	206	54 (26.21%)	152 (73.79%)
	Third Quintile	336	90 (26.79%)	246 (73.21%)	244	60 (24.59%)	184 (75.41%)
	Fourth Quintile	509	140 (27.50%)	369 (72.50%)	206	72 (34.95%)	134 (65.05%)
	Fifth Quintile	1203	368 (30.59%)	835 (69.41%)	182	64 (35.16%)	118 (64.84%)
Locale	City	708	211 (29.80%)	497 (70.20%)	315	98 (31.11%)	217 (68.89%)
	Suburb	884	240 (27.15%)	644 (72.85%)	393	108 (27.48%)	285 (72.52%)
	Town	291	104 (35.74%)	187 (64.26%)	101	33 (32.67%)	68 (67.33%)
	Rural	514	155 (30.16%)	359 (69.84%)	225	60 (26.67%)	165 (73.33%)

Note. The term Indigenous refers to students who are American Indian, Alaska Native, Native Hawaiian, and Pacific Islander.

Figure 1 shows the random forest feature importance bar graph for high-achieving students. The feature importance scores for gender, race, socioeconomic status were 1.77%, 4.53%, and 4.8%, respectively, indicating that among these three demographic variables socioeconomic status was the most important.

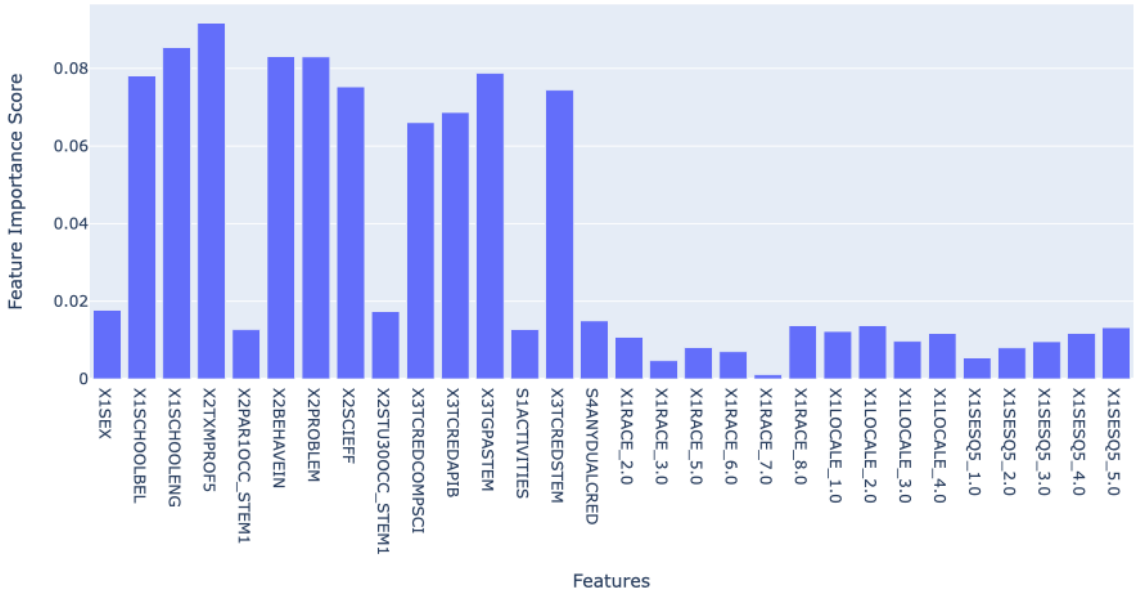

Figure 1. Feature Importance Bar Graph for High-Achieving Students Based on SMOTE

Note. X1SCHOOLBEL = School Belonging, X1SCHOOLENG = School Engagement, X2TXMPROF5 = Math Proficiency, X2PAR1OCC_STEM1 = Parent Occupation, X2BEHAVEIN = School Motivation, X2PROBLEM = School Problems, X2SCIEFF = Science Self-Efficacy, X2STU30OCC_STEM1 = Expected Occupation at 30, X3TCREDCOMPSCI = Credits in Computer Science, X3TCREDAPIB = Credits in AP/IB combined, X3TGPASTEM = GPA in STEM, S1ACTIVITIES = Activity Attendance, X3TCREDSTEM = Credits in STEM, S4ANYDUALCRED = Dual Credits, X1RACE_2.0 = Asian, X1RACE_3.0 = Black, X1RACE_5.0 = Hispanic, X1RACE_6.0 = Multiracial, X1RACE_7.0 = Indigenous, X1RACE_8.0 = White, X1LOCALE_1.0 = City, X1LOCALE_2.0 = Suburb, X1LOCALE_3.0 = Town, X1LOCALE_4.0 = Rural, X1SESQ5_1.0 = First quintile (lowest), X1SESQ5_2.0 = Second quintile, X1SESQ5_3.0 = Third quintile, X1SESQ5_4.0 = Fourth quintile, X1SESQ5_5.0 = Fifth quintile (highest)

An examination of each demographic variable reveals that the feature importance score of race-related dummy variables was smaller than that of other variables included in the study. The highest feature importance score was found for White students and the lowest for Indigenous students. Furthermore, socioeconomic status also has small feature importance scores; however,

the feature importance scores increase from the first to the fifth quantile. This result indicates that students from higher socioeconomic backgrounds are more likely to persist in STEM. Examination of the importance of locale information importance indicated that suburban areas have higher feature importance than other areas.

Figure 2 shows the random forest feature importance bar graph for non-high-achieving students. The feature importance scores for gender, race, and socioeconomic status were 2.5%, 5.89%, and 6.37%, respectively, indicating that, similar to results for high-achieving students, economic status was the most important among these three demographic variables. However, non-high-achieving students had a higher importance score for socioeconomic status than high-achieving students.

An examination of each demographic variable reveals that the feature importance of race was higher for Hispanic and White students than for other groups, and it was lowest for Indigenous students. Overall, the feature importance scores of race related variables were much lower than those of other variables included in the study. Furthermore, the third quintile had the highest feature importance score regarding socioeconomic status while the first quintile had the lowest score. Among the four locales, students from suburban areas had the highest feature importance score, implying that they were more likely to persist in STEM.

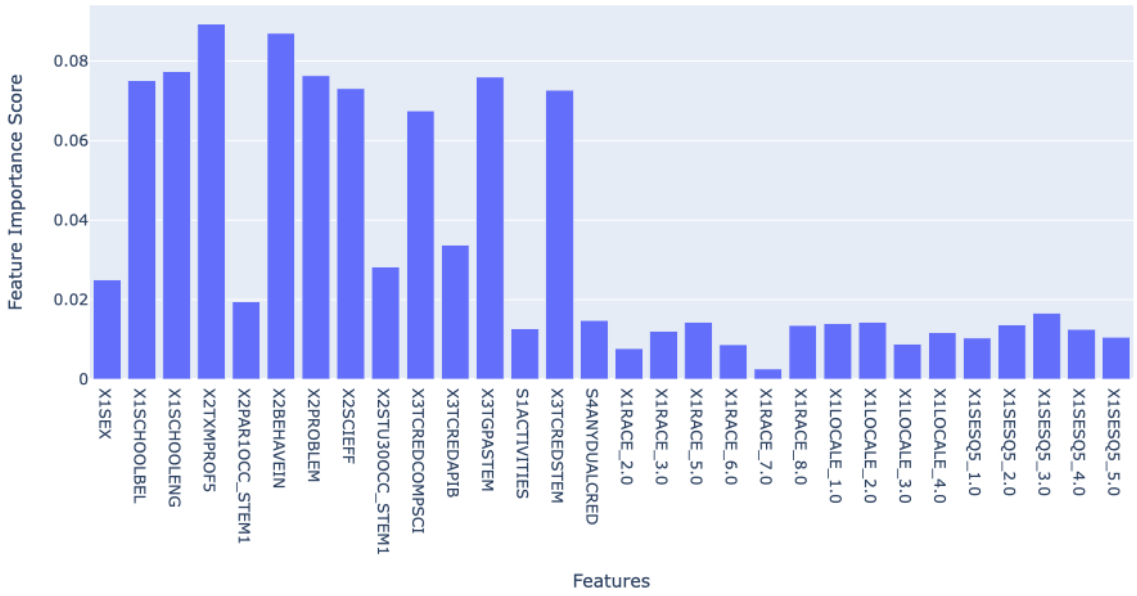

Figure 2. Feature Importance Bar Graph for Non-High-Achieving Students Based on SMOTE

Note. X1SCHOOLBEL = School Belonging, X1SCHOOLENG = School Engagement, X2TXMPROF5 = Math Proficiency, X2PAR1OCC_STEM1 = Parent Occupation, X2BEHAVEIN = School Motivation, X2PROBLEM = School Problems, X2SCIEFF = Science Self-Efficacy, X2STU30OCC_STEM1 = Expected Occupation at 30, X3TCREDCOMPSCI = Credits in Computer Science, X3TCREDAPIB = Credits in AP/IB combined, X3TGPASTEM = GPA in STEM, S1ACTIVITIES = Activity Attendance, X3TCREDSTEM = Credits in STEM, S4ANYDUALCRED = Dual Credits, X1RACE_2.0 = Asian, X1RACE_3.0 = Black, X1RACE_5.0 = Hispanic, X1RACE_6.0 = Multiracial, X1RACE_7.0 = Indigenous, X1RACE_8.0 = White, X1LOCALE_1.0 = City, X1LOCALE_2.0 = Suburb, X1LOCALE_3.0 = Town, X1LOCALE_4.0 = Rural, X1SESQ5_1.0 = First quintile (lowest), X1SESQ5_2.0 = Second quintile, X1SESQ5_3.0 = Third quintile, X1SESQ5_4.0 = Fourth quintile, X1SESQ5_5.0 = Fifth quintile (highest)

Research Question 5: Which machine learning techniques can be used to identify variables influencing the early postsecondary persistence of high-achieving and non-high-achieving students in terms of classification models?

In this study, I used random forest and artificial neural networks as machine learning models. In addition, I used original and augmented (SMOTE) student data to obtain the results. The descriptive statistics of each variable are shown in Tables 8 through 11. Furthermore, I present results for high-achieving and non-high-achieving students separately, including random forest and neural network results with original and SMOTE datasets to better understand how variables influence the early postsecondary STEM persistence of these groups.

Table 8 presents descriptive statistics that summarize the central tendency, dispersion, and shape of the imbalanced dataset distribution for high-achieving students, including the count, mean, standard deviation, and dispersion of each feature.

Table 8. Descriptive Statistics That Summarize the Central Tendency, Dispersion and Shape of Imbalanced Dataset Distribution for High-Achieving Students (GPA >= 3.5)

Feature	count	mean	std	min	25%	50%	75%	max
Gender	2397	0.411348	0.492181	0	0.000000	0.000000	1.000000	1
School Belonging	2397	0.782007	0.149772	0	0.694757	0.780899	0.908240	1
School Engagement	2397	0.815607	0.164994	0	0.729560	0.849057	0.937107	1
Math Proficiency	2397	0.533440	0.414498	0	0.055286	0.661465	0.959595	1
Parent Occupation	2397	0.203171	0.402442	0	0.000000	0.000000	0.000000	1
School Motivation	2397	0.867598	0.113251	0	0.823633	0.892416	0.943563	1
School Problems	2397	0.368154	0.214473	0	0.210884	0.361451	0.494331	1
Science Self-Efficacy	2397	0.694052	0.216390	0	0.615572	0.666667	0.863747	1
Expected Occupation at 30	2397	0.666667	0.471503	0	0.000000	1.000000	1.000000	1
Credits in Computer Science	2397	0.105924	0.146944	0	0.000000	0.100000	0.200000	1
Credits in AP/IB combined	2397	0.250088	0.250617	0	0.000000	0.153846	0.384615	1
GPA in STEM	2397	0.883187	0.097399	0	0.875000	0.875000	1.000000	1
Activity Attendance	2397	0.232791	0.422698	0	0.000000	0.000000	0.000000	1
Credits in STEM	2397	0.571131	0.133682	0	0.483871	0.548387	0.645161	1
Dual Credits	2397	0.372132	0.483474	0	0.000000	0.000000	1.000000	1
Asian	2397	0.179808	0.384108	0	0.000000	0.000000	0.000000	1
Black	2397	0.041302	0.199028	0	0.000000	0.000000	0.000000	1
Hispanic	2397	0.092199	0.289366	0	0.000000	0.000000	0.000000	1
Multiracial	2397	0.070505	0.256049	0	0.000000	0.000000	0.000000	1
Indigenous	2397	0.006258	0.078875	0	0.000000	0.000000	0.000000	1
White	2397	0.609929	0.487868	0	0.000000	1.000000	1.000000	1
City	2397	0.295369	0.456304	0	0.000000	0.000000	1.000000	1
Suburb	2397	0.368794	0.482579	0	0.000000	0.000000	1.000000	1
Town	2397	0.121402	0.326662	0	0.000000	0.000000	0.000000	1
Rural	2397	0.214435	0.410515	0	0.000000	0.000000	0.000000	1
First quintile (lowest)	2397	0.055903	0.229783	0	0.000000	0.000000	0.000000	1
Second quintile	2397	0.089695	0.285805	0	0.000000	0.000000	0.000000	1
Third quintile	2397	0.140175	0.347241	0	0.000000	0.000000	0.000000	1
Fourth quintile	2397	0.212349	0.409056	0	0.000000	0.000000	0.000000	1
Fifth quintile (highest)	2397	0.501877	0.500101	0	0.000000	1.000000	1.000000	1
Persistent	2397	0.703796	0.456677	0	0.000000	1.000000	1.000000	1

As shown in Table 8, school motivation and GPA in STEM have higher means, showing that students' school motivation and GPA in STEM are higher compared to other features for high-achieving students. However, math proficiency variable had the highest score (0.959595) at the 75%, showing that 25% of students' math proficiency score are close the highest score. Another point to be addressed is that among the means of racial groups, the highest was White and the lowest was Indigenous, showing that a greater number of White students are high-achieving than of other groups.

Table 9 presents descriptive statistics that summarize the central tendency, dispersion, and shape of the SMOTE dataset distribution for high-achieving students, including the count, mean, standard deviation, and dispersion of each feature.

Table 9. Descriptive Statistics That Summarize the Central Tendency, Dispersion and Shape of the SMOTE Dataset Distribution for High-Achieving Students (GPA >= 3.5)

feature	count	mean	std	min	25%	50%	75%	max
Gender	3374	0.375519	0.484328	0	0.000000	0.000000	1.000000	1
School Belonging	3374	0.779352	0.143844	0	0.691011	0.776626	0.889513	1
School Engagement	3374	0.814651	0.156425	0	0.729560	0.840671	0.920335	1
Math Proficiency	3374	0.521610	0.407157	0	0.062984	0.617598	0.945618	1
Parent Occupation	3374	0.179609	0.383918	0	0.000000	0.000000	0.000000	1
School Motivation	3374	0.865185	0.109698	0	0.823633	0.888598	0.940035	1
School Problems	3374	0.369354	0.206715	0	0.222222	0.365592	0.492063	1
Science Self-Efficacy	3374	0.688932	0.209042	0	0.605147	0.666667	0.844282	1
Expected Occupation at 30	3374	0.640486	0.479929	0	0.000000	1.000000	1.000000	1
Credits in Computer Science	3374	0.103865	0.138598	0	0.000000	0.078398	0.200000	1
Credits in AP/IB combined	3374	0.245817	0.240949	0	0.053942	0.162562	0.384615	1
GPA in STEM	3374	0.881431	0.092900	0	0.840652	0.875000	1.000000	1
Activity Attendance	3374	0.201541	0.401211	0	0.000000	0.000000	0.000000	1
Credits in STEM	3374	0.569426	0.126896	0	0.483871	0.548387	0.637864	1
Dual Credits	3374	0.348548	0.476581	0	0.000000	0.000000	1.000000	1
Asian	3374	0.175163	0.380163	0	0.000000	0.000000	0.000000	1
Black	3374	0.035862	0.185975	0	0.000000	0.000000	0.000000	1
Hispanic	3374	0.082395	0.275006	0	0.000000	0.000000	0.000000	1
Multiracial	3374	0.069057	0.253589	0	0.000000	0.000000	0.000000	1
Indigenous	3374	0.004446	0.066538	0	0.000000	0.000000	0.000000	1
White	3374	0.600771	0.489813	0	0.000000	1.000000	1.000000	1
City	3374	0.295199	0.456200	0	0.000000	0.000000	1.000000	1
Suburb	3374	0.354475	0.478425	0	0.000000	0.000000	1.000000	1
Town	3374	0.122407	0.327803	0	0.000000	0.000000	0.000000	1
Rural	3374	0.213693	0.409973	0	0.000000	0.000000	0.000000	1
First quintile (lowest)	3374	0.048014	0.213828	0	0.000000	0.000000	0.000000	1
Second quintile	3374	0.086544	0.281208	0	0.000000	0.000000	0.000000	1
Third quintile	3374	0.130705	0.337128	0	0.000000	0.000000	0.000000	1
Fourth quintile	3374	0.203023	0.402309	0	0.000000	0.000000	0.000000	1
Fifth quintile (highest)	3374	0.505039	0.500049	0	0.000000	1.000000	1.000000	1
Persistent	3374	0.500000	0.500074	0	0.000000	0.500000	1.000000	1

Table 9 shows that the descriptive statistics of the SMOTE dataset for high-achieving students are similar to those of the imbalanced dataset. In this dataset, school motivation and GPA in STEM have higher means than other features for high-achieving students. Also, math proficiency still had the highest score (0.945618) at 75%, showing that 25% of the students' math proficiency scores are close the highest score. Also, the race variable had a similar pattern in the SMOTE dataset. The mean score of White was the highest while Indigenous was the lowest, showing that a greater number of White students are high-achieving based on high school GPA.

Table 10 presents descriptive statistics that summarize the central tendency, dispersion, and shape of imbalanced dataset distribution for non-high-achieving students, including the count, mean, standard deviation, and dispersion of each feature.

Table 10. Descriptive Statistics That Summarize the Central Tendency, Dispersion and Shape of Imbalanced Dataset Distribution for Non-High-Achieving Students (GPA <= 2.5)

feature	count	mean	std	min	25%	50%	75%	max
Gender	1034	0.488395	0.500107	0	0.000000	0.000000	1.000000	1
School Belonging	1034	0.699844	0.175415	0	0.614095	0.689524	0.824762	1
School Engagement	1034	0.694545	0.194238	0	0.587002	0.712998	0.840671	1
Math Proficiency	1034	0.100000	0.239005	0	0.001033	0.006408	0.036663	1
Parent Occupation	1034	0.150870	0.358096	0	0.000000	0.000000	0.000000	1
School Motivation	1034	0.812063	0.129217	0	0.754745	0.837178	0.902190	1
School Problems	1034	0.436793	0.193805	0	0.322176	0.451883	0.550941	1
Science Self-Efficacy	1034	0.628858	0.219145	0	0.525547	0.666667	0.691241	1
Expected Occupation at 30	1034	0.500967	0.500241	0	0.000000	1.000000	1.000000	1
Credits in Computer Science	1034	0.101838	0.147868	0	0.000000	0.000000	0.200000	1
Credits in AP/IB combined	1034	0.046608	0.119815	0	0.000000	0.000000	0.000000	1
GPA in STEM	1034	0.439329	0.146849	0	0.333333	0.466667	0.600000	1
Activity Attendance	1034	0.111219	0.314555	0	0.000000	0.000000	0.000000	1
Credits in STEM	1034	0.475369	0.152073	0	0.406250	0.500000	0.562500	1
Dual Credits	1034	0.165377	0.371700	0	0.000000	0.000000	0.000000	1
Asian	1034	0.069632	0.254650	0	0.000000	0.000000	0.000000	1
Black	1034	0.182785	0.386677	0	0.000000	0.000000	0.000000	1
Hispanic	1034	0.205996	0.404623	0	0.000000	0.000000	0.000000	1
Multiracial	1034	0.094778	0.293049	0	0.000000	0.000000	0.000000	1
Indigenous	1034	0.018375	0.134369	0	0.000000	0.000000	0.000000	1
White	1034	0.428433	0.495091	0	0.000000	0.000000	1.000000	1
City	1034	0.304642	0.460478	0	0.000000	0.000000	1.000000	1
Suburb	1034	0.380077	0.485640	0	0.000000	0.000000	1.000000	1
Town	1034	0.097679	0.297024	0	0.000000	0.000000	0.000000	1
Rural	1034	0.217602	0.412815	0	0.000000	0.000000	0.000000	1
First quintile (lowest)	1034	0.189555	0.392139	0	0.000000	0.000000	0.000000	1
Second quintile	1034	0.199226	0.399612	0	0.000000	0.000000	0.000000	1
Third quintile	1034	0.235977	0.424813	0	0.000000	0.000000	0.000000	1
Fourth quintile	1034	0.199226	0.399612	0	0.000000	0.000000	0.000000	1
Fifth quintile (highest)	1034	0.176015	0.381018	0	0.000000	0.000000	0.000000	1
Persistent	1034	0.710832	0.453596	0	0.000000	1.000000	1.000000	1

Descriptive statistics of the imbalanced dataset for non-high-achieving students show that the mean of each variable is lower for them than for high-achieving students. Despite their low STEM GPAs, however, the mean of their school motivation was 0.812063, suggesting their high motivation to persist. Also, their math proficiency mean is 0.100000, showing that non-high-achieving students have lower scores in math. Furthermore, the mean value differences among race variables demonstrated that White students had the highest value, and Indigenous students had the lowest, showing that a greater number of White students are non-high-achieving. However, the gap between these values is lower compared to that of high-achieving students.

Table 11 represents descriptive statistics that summarize the central tendency, dispersion, and shape of the SMOTE dataset distribution for non-high-achieving students, including the count, mean, standard deviation, and dispersion of each feature.

Table 11. Descriptive Statistics That Summarize the Central Tendency, Dispersion and Shape of SMOTE Dataset Distribution for Non-High-Achieving Students (GPA <= 2.5)

feature	count	mean	std	min	25%	50%	75%	max
Gender	1470	0.475549	0.489322	0	0.000000	0.066230	1.000000	1
School Belonging	1470	0.694782	0.166221	0	0.611809	0.688440	0.801910	1
School Engagement	1470	0.699221	0.183490	0	0.588077	0.714890	0.840670	1
Math Proficiency	1470	0.096171	0.225838	0	0.001757	0.007650	0.039780	1
Parent Occupation	1470	0.138578	0.337678	0	0.000000	0.000000	0.000000	1
School Motivation	1470	0.813022	0.122653	0	0.760584	0.837960	0.899270	1
School Problems	1470	0.431873	0.187461	0	0.320084	0.437590	0.546030	1
Science Self-Efficacy	1470	0.629511	0.207360	0	0.530414	0.661110	0.703160	1
Expected Occupation at 30	1470	0.482980	0.490418	0	0.000000	0.192660	1.000000	1
Credits in Computer Science	1470	0.094990	0.136088	0	0.000000	0.012440	0.195240	1
Credits in AP/IB combined	1470	0.043784	0.112262	0	0.000000	0.000000	0.020020	1
GPA in STEM	1470	0.442336	0.138598	0	0.333333	0.466670	0.545390	1
Activity Attendance	1470	0.099863	0.293333	0	0.000000	0.000000	0.000000	1
Credits in STEM	1470	0.476692	0.142484	0	0.406250	0.500000	0.562500	1
Dual Credits	1470	0.163562	0.361707	0	0.000000	0.000000	0.000000	1
Asian	1470	0.053741	0.225584	0	0.000000	0.000000	0.000000	1
Black	1470	0.165986	0.372195	0	0.000000	0.000000	0.000000	1
Hispanic	1470	0.178231	0.382838	0	0.000000	0.000000	0.000000	1
Multiracial	1470	0.079592	0.270752	0	0.000000	0.000000	0.000000	1
Indigenous	1470	0.016327	0.126771	0	0.000000	0.000000	0.000000	1
White	1470	0.453741	0.498025	0	0.000000	0.000000	1.000000	1
City	1470	0.297279	0.457216	0	0.000000	0.000000	1.000000	1
Suburb	1470	0.372789	0.483711	0	0.000000	0.000000	1.000000	1
Town	1470	0.084354	0.278012	0	0.000000	0.000000	0.000000	1
Rural	1470	0.207483	0.405642	0	0.000000	0.000000	0.000000	1
First quintile (lowest)	1470	0.172109	0.377603	0	0.000000	0.000000	0.000000	1
Second quintile	1470	0.187075	0.390104	0	0.000000	0.000000	0.000000	1
Third quintile	1470	0.212925	0.409514	0	0.000000	0.000000	0.000000	1
Fourth quintile	1470	0.212925	0.409514	0	0.000000	0.000000	0.000000	1
Fifth quintile (highest)	1470	0.182313	0.386234	0	0.000000	0.000000	0.000000	1
Persistent	1470	0.500000	0.500170	0	0.000000	0.500000	1.000000	1

Table 11 shows that the descriptive statistics of the SMOTE dataset for non-high-achieving students were similar to those of the imbalanced dataset. For example, the mean value of school motivation (0.813022) is similar to that of the imbalanced dataset results. Also, GPA in STEM is 0.442336, similar to that of non-high-achieving students in the imbalanced dataset results but much lower than high-achieving students' results. Thus, the imbalanced and SMOTE dataset provides similar results for non-high-achieving students but shows different trends from those of high-achieving students.

Machine Learning Results for High-Achieving Students

In this study, 2,397 STEM students had GPAs of 3.5 and higher. Of these students, 1,687 (~70%) continued their STEM education, while the remaining 710 (~30%) did not. I performed four different analyses for high-achieving students that yielded (1) results with the original dataset including random forest and artificial neural network, and (2) results with the SMOTE dataset including random forest and artificial neural network. The test results (e.g., accuracy, sensitivity, and specificity) for the machine learning algorithms for high-achieving students in the imbalanced and the SMOTE datasets are presented in Table 12.

Table 12. Machine Learning Algorithms Test Results for High-Achieving Students (GPA >= 3.5)

Dataset	Method	Sample	Negative Data	Positive Data	Training Set Size	Test Set Size	Neg. Test Size	Pos. Test Size	Acc	Sens	Spec
Imbalanced	Random Forest	(2397, 30)	(710, 30)	(1687, 30)	(1917, 30)	(480, 30)	158	322	0.66	0	0.99
Imbalanced	ANN	(2397, 30)	(710, 30)	(1687, 30)	(1917, 30)	(480, 30)	158	322	0.88	0.75	0.94
SMOTE	Random Forest	(3374, 30)	(1687, 30)	(1687, 30)	(2699, 30)	(675, 30)	342	333	0.82	0.75	0.88
SMOTE	ANN	(3374, 30)	(1687, 30)	(1687, 30)	(2699, 30)	(675, 30)	342	333	0.92	0.94	0.89

Results with Original (Imbalanced) Dataset for High-Achieving Students

In this study, the original data were imbalanced in that the minority class (non-persistent students) included fewer students than the majority class. Random forest and artificial neural network results based on the original dataset are shown in Table 12.

The accuracy of the random forest model (see Table 12) was 0.66, and its specificity and sensitivity were 0.99 and 0, respectively. Compared to random forest model, the artificial neural network method provided much better results, as accuracy, sensitivity, and specificity were 0.88, 0.75, and 0.94, respectively. However, the sensitivity of artificial neural network was less than the specificity, implying that the model's ability to predict the minority class's (non-persistent students') persistence could still be improved.

That the majority class (persistent students) included more students than the minority class caused the imbalanced learning problem with the use of the original dataset. Therefore, in this study, the SMOTE technique was used to avoid the imbalanced learning problem in further analyses.

Results with Augmented (SMOTE) Dataset for High-Achieving Students

To increase the performance of random forest and artificial neural network, the SMOTE imbalanced learning technique was used in this study to increase the accuracy of both the random forest and artificial neural network models. Based on the random forest model (see Table 12), the accuracy was 0.82, while the model's specificity and sensitivity were 0.75 and 0.88 respectively. Furthermore, the artificial neural network's result had an accuracy of 0.92 (see Table 12), while the model's sensitivity and specificity were 0.94 and 0.89 respectively.

Machine Learning Results for Non-High-Achieving Students

In this study, there were 1,034 STEM students with a GPA of 2.5 and lower, of whom 735 (71%) continued their STEM education, while the remaining 299 (29%) did not. I performed four different analyses for non-high-achieving students: random forest and artificial neural network results with the original and SMOTE datasets. Results for non-high-achieving students are presented in Table 13.

Table 13. Machine Learning Algorithms Test Results for Non-High-Achieving Students (GPA <= 2.5)

Dataset	Method	Sample	Negative Data	Positive Data	Training Set Size	Test Set Size	Neg. Test Size	Pos. Test Size	Acc	Sens	Spec
Imbalanced	Random Forest	(1034, 30)	(299, 30)	(735, 30)	(827, 30)	(207, 30)	53	154	0.74	0.04	0.99
Imbalanced	ANN	(1034, 30)	(299, 30)	(735, 30)	(828, 30)	(206, 30)	53	154	0.95	0.92	0.96
SMOTE	Random Forest	(1470, 30)	(735, 30)	(735, 30)	(2699, 30)	(675, 30)	149	145	0.85	0.8	0.91
SMOTE	ANN	(1470, 30)	(735, 30)	(735, 30)	(2699, 30)	(675, 30)	149	145	0.96	0.97	0.96

Results with Original (Imbalanced) Dataset for Non-High-Achieving Students

In this study, the original data for non-high-achieving students was imbalanced in that the minority class (non-persistent students) was smaller than the majority class. Random forest and artificial neural network results based on the original dataset are given in Table 13. The accuracy of the random forest model (see Table 13) was 0.74, while the model's specificity and sensitivity were 0.99 and 0.04, respectively. Based on the original data results, the artificial neural network results for non-high-achieving students were 0.95 for accuracy and 0.92 for sensitivity (see Table 13). The SMOTE technique was used to avoid the imbalanced learning problem in further analyses.

Results with Augmented (SMOTE) Dataset for Non-High-Achieving Students

To increase the performance of the random forest and artificial neural network models for non-high-achieving students, the SMOTE technique was used in this study. SMOTE results have shown that using this common imbalanced learning technique can increase a model's accuracy. The accuracy of the random forest model for non-high-achieving students (see Table 13), was 0.85, and its specificity and sensitivity were 0.91 and 0.8 respectively. In comparison, the accuracy of the artificial neural network model for this group was 0.96 (see Table 13) and its sensitivity and specificity were 0.97 and 0.96 respectively. These results showed the neural network model could predict non-high-achieving students' persistence with a high level of accuracy. Moreover, it performed well with both non-persistent and persistent groups.

Receiver Operating Characteristics (ROC) Curve

As the ROC curve is one of the evaluation measures of the study measures to better identify accuracy using various thresholds, I present the performance analysis of machine learning models

for original (imbalanced) and augmented synthetic student (SMOTE) data using ROC curves in Figures 3 through 6. An ROC curve is plotted by computing the true positive rate and the false positive rate for different threshold values at the probability output of machine learning models. True positive rate represents the instances of a positive label (here, "persistent") the model correctly identifies out of all the possible instances. In contrast, false positive rate is the probability of a false report when an instance of a negative label ("non-persistent") is categorized as positive ("persistent"). Because true positive rate and false positive rate values change according to selected threshold changes, any threshold value can be selected to lower the false report rate or increase the detection accuracy based on the system's specific needs.

In this study, machine learning models were used to classify student data into two different classes based on the highest probability using the last layer's output. The student data were classified as persistent in STEM education if the probability of persistence was higher than 0.5. However, the best performance might not be obtained using equal probability. Thus, I used ROC curves to conduct a performance analysis to examine where accuracy was plotted based on various thresholds. It should be noted that in this research sensitivity is the most important factor because its result shows the success rates of the detection of persistency in STEM education.

As shown in Figures 4 and 6, in the ROC curves analyses of the augmented synthetic student (SMOTE) data, marked with a solid orange line, show better results were more accurate than those of the analyses of the original (imbalanced) student data, shown in Figures 3 and 5. When the model was trained with a dataset that included additional data, in particular an increased amount of augmented non-persistent student data, it performed more accurately than with the imbalanced dataset. Furthermore, as two commonly used machine learning algorithms for non-linearly distributed data, random forest and artificial neural network produce more accurate results

with augmented data, as shown in Tables 12 and 13 above, indicating that a model's accuracy depends on the size of samples. Accuracy increases as the model complexity increases along with the usage of non-linear based predictive machine learning models as presented in ROC curves and test results as shown in Tables 12 and 13.

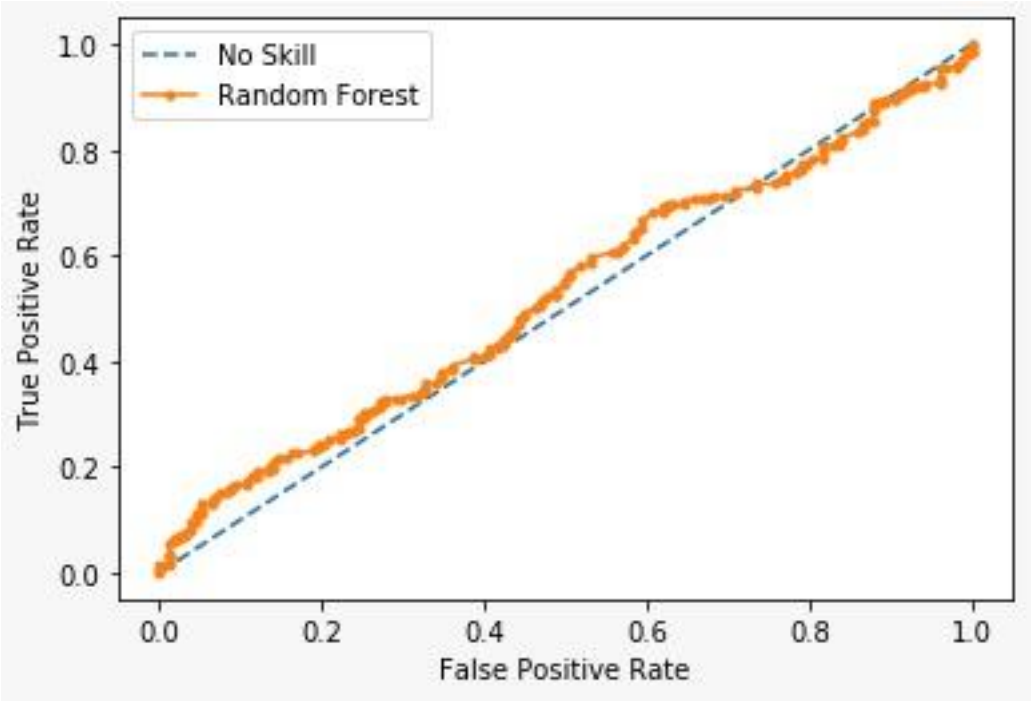

Figure 3. ROC Curve Results for High-Achieving Students with Imbalanced Dataset

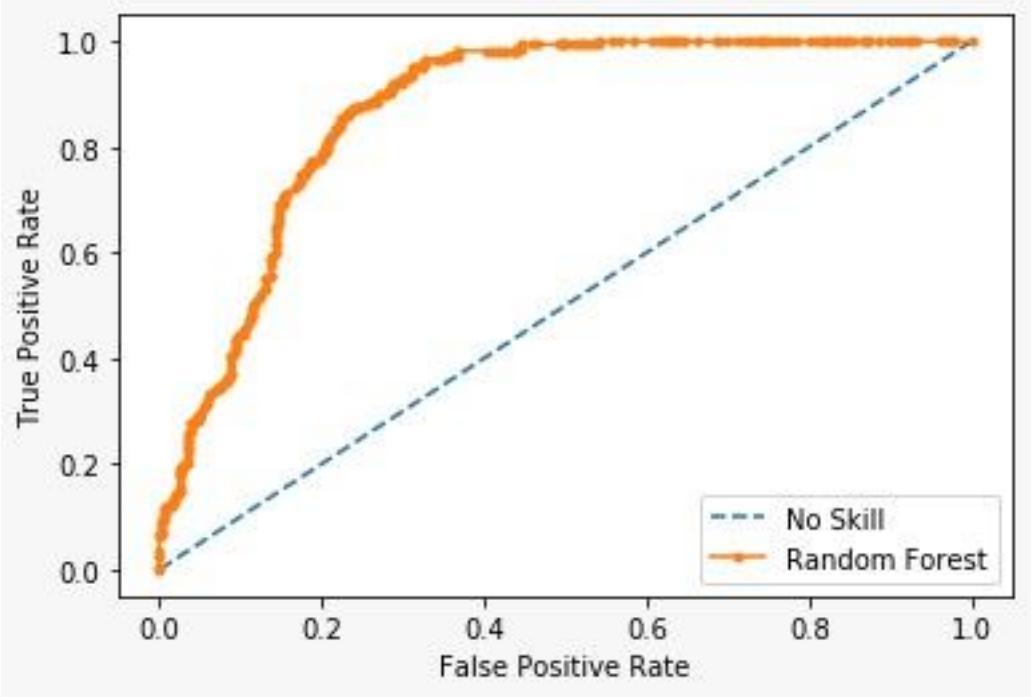

Figure 4. ROC Curve Results for High-Achieving Students with SMOTE Dataset

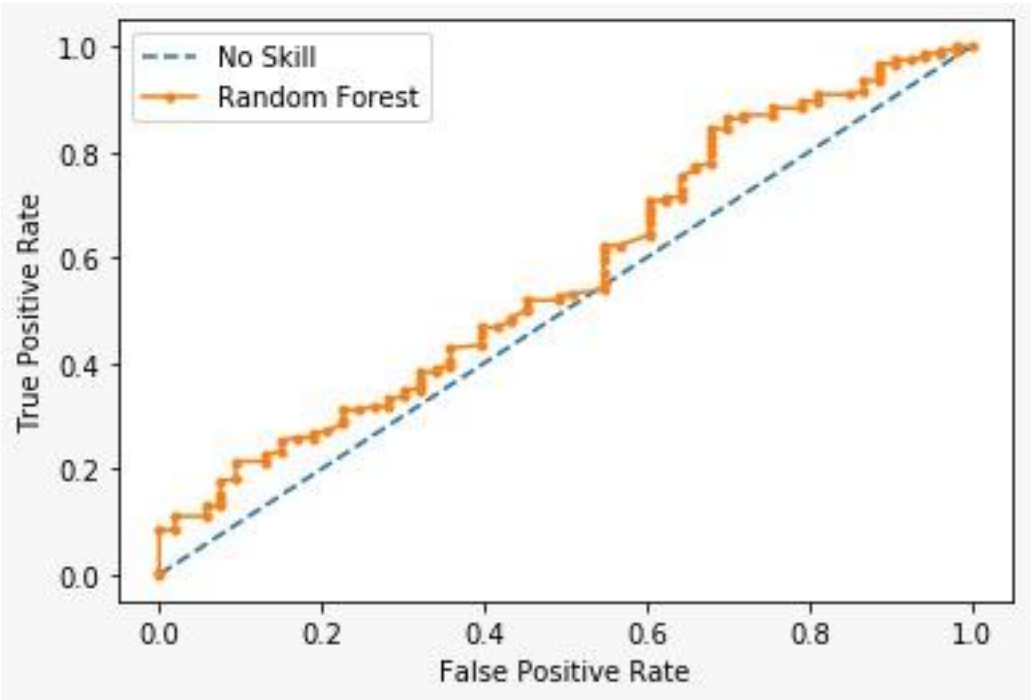

Figure 5. ROC Curve Results for Non-High-Achieving Students with Imbalanced Dataset

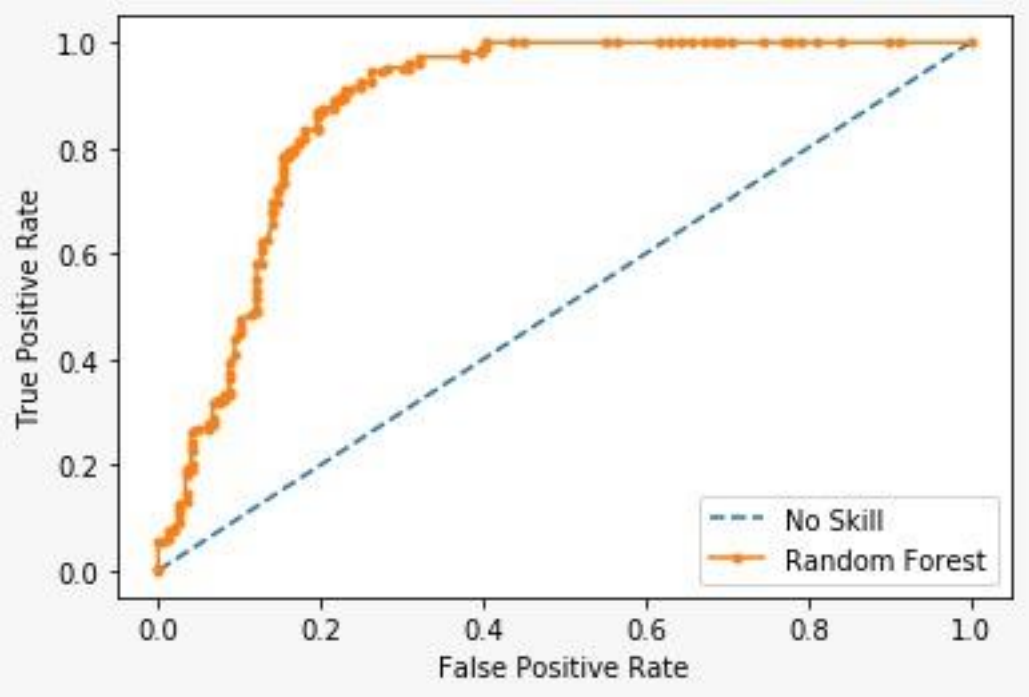

Figure 6. ROC Curve Results for Non-High-Achieving Students with SMOTE Dataset

Based on the ROC Curve results for both high-achieving and non-achieving students, the SMOTE dataset performed more accurately than the imbalanced dataset (see figures 4 and 6).

CHAPTER 5 DISCUSSION

In this dissertation study, I examined factors affecting high-achieving and non-high-achieving students' early postsecondary STEM persistence using machine learning techniques and methods. The following discussion of my findings is organized according to the research questions.

Research Question 1: What percentages of high-achieving and non-high-achieving students persist in postsecondary STEM majors?

Previous research has mostly focused on general student persistence (Aulck et al., 2017; Chen et al., 2018; Cardona et al., 2020), and few studies have specifically addressed high-achieving and non-high-achieving students' persistence in STEM fields in college. To determine high-achieving and non-high-achieving students' persistence in STEM, a nationally representative dataset, the HSLS: 2009, was used in the study. The results demonstrated that high-achieving and non-high-achieving students' persistence levels were similar; however, the sample was imbalanced in that it comprised more than twice as many high-achieving as non-high-achieving students in STEM fields. Based on the literature (Heilbronner, 2011; Steenbergen-Hu & Olszewski-Kubilius, 2017), high-achieving students are more likely to have the potential to pursue a STEM field. The study results supported the literature showing that there were more high-achieving students in STEM fields. This can be a result of having more opportunities in STEM, such as taking advanced STEM courses.

The results of this study are in line with literature showing that nearly a third of students with STEM majors drop out of the field prior to graduation (Green & Anderson, 2018; NCES, 2018), while approximately 70% of both high-achieving and non-high-achieving students persist

in STEM in college. Thus, these results provide confirmatory evidence that persistence in STEM fields is problematic regardless of academic achievement.

In addition, it is important to note that according to National Center for Education Statistics (NCES, 2017) report, 70% of first-time postsecondary students persisted at their institutions after three years of initial enrollment. The study results were similar to persistence in general in postsecondary education institutions.

Furthermore, the study results revealed that high-achieving and non-high-achieving students had a similar persistence rates after entering a STEM field. This study provides evidence that entering a STEM field can be a turning point in the lives of non-high-achieving students regarding STEM persistence. Therefore, educators should also support non-high-achieving students' entrance into STEM fields as they likewise have the potential to persist in STEM.

Research Question 2: What variables affect high-achieving and non-high-achieving students' persistence in postsecondary STEM majors?

Based on my literature review and statistical selection techniques, the variables selected for this study were gender, school belonging, school engagement, math proficiency level, parent occupation, school motivation, school problems, science self-efficacy, expected occupation at age 30, credits earned in computer sciences, GPA for STEM courses, student activities, credits earned in STEM, credits earned in AP/IB combined, dual credits, race, socioeconomic status, and urbanicity. The results align with previous studies showing the importance of demographics (Holmes et al., 2018; Turner et al., 2019), cognitive aspects (Watkins & Mazur, 2013; Nicholls et al., 2010) and non-cognitive aspects (Aryee 2017; Dimer & Li, 2012, Lent et al., 1994, 2000; Heilbronner 2011) while adding to the literature regarding significant non-cognitive variables. Most studies involving machine learning have focused on demographics and college level

cognitive factors (Thammasiri et al., 2014). This study expanded machine learning research by providing evidence that even high school non-cognitive variables can be used to predict students' early postsecondary STEM persistence. In general, the results demonstrate the importance of supporting students' social, emotional, and academic development while in high school.

According to Mendez et al. (2008), predicting student persistence in STEM is not an easy task and requires advanced methods; however, the authors argued that machine learning could deal with complex relationships between and among variables. In this study, students' STEM persistence was predicted using random forest and artificial neural network models, which provided high accuracy with selected variables. Most previous machine learning studies in education have focused on single feature importance predictors (Aulck et al., 2016; Baranyi et al., 2020). However, in this study, I used the random forest feature importance technique to determine the importance of each of several features and how they affect the persistence of students in STEM overall. This technique contributes to educational research in education by demonstrating and comparing the importance of variables included in the final model.

Research Question 3: Which variables most significantly influence the early postsecondary persistence of high-achieving and non-high-achieving students?

According to Sage et al. (2018), academic performance is a strong predictor that supports student persistence in college. Previous research has indicated that STEM related courses (Mendez et al., 2008; Tyson et al., 2007; Wang, 2013), scores (Chimka et al., 2007; French et al., 2005; Min et al., 2011; Nicholls et al., 2007; Watkins & Mazur, 2013; Zhang et al., 2004), and credits (Chen & Soldner, 2013) positively influence student persistence; this study confirms previous results showing that math scores, credits earned in computer sciences, GPA in STEM courses, credits earned in STEM and AP/IB courses significantly predict student persistence in STEM for both

high-achieving and non-high-achieving students. Furthermore, this research expands the literature by showing that credits taken in AP/IB courses can be used to predict student persistence. Additionally, credits taken in AP/IB courses were two times more important for high-achieving students, showing the greater importance of these courses for STEM persistence. This might be a result of educators' recommendations of taking these courses, especially for high-achieving students. Additionally, in general, high-achieving students might link these courses with potential success. However, AP/IB courses are also important for non-high-achieving students. Thus, educators should encourage both high-achieving and non-high-achieving students to take AP/IB courses as they lead to persistence in STEM. Besides, schools can offer more AP/IB courses, as it is an important factor for STEM persistence.

The results of this study provide evidence that non-cognitive factors (e.g., school belonging, science self-efficacy, school motivation, and school engagement) significantly influence both high-achieving and non-high-achieving students' persistence in STEM. These results are consistent with previous research demonstrating the importance of non-cognitive factors (Aryee, 2017; Nugent et al., 2015). Furthermore, I included variables related to non-cognitive factors that have not been widely investigated in a single study to determine persistence in STEM for both high-achieving and non-high-achieving students.

Some differences existed regarding feature importance scores among non-cognitive variables. For instance, compared to high-achieving students, non-high-achieving students' feature importance score of school motivation was higher, indicating the greater importance of school motivation for non-high-achieving students. However, motivation was among the most important variables for both high-achieving students and non-high-achieving students, indicating importance of the school motivation in general.

In addition, based on the literature, self-efficacy is positively associated with student achievement and motivation (Nugent et al., 2015). In this study, science self-efficacy was among the most important variables for high-achieving students and non-high-achieving students. Overall, the results indicate that even though slight differences exist in feature importance scores, researchers should investigate non-cognitive variables as they are also crucial to support students' persistence in STEM.

Research Question 4: To what extent do high-achieving and non-high-achieving students' demographics (e.g., gender, ethnicity, socioeconomic status) affect their early postsecondary STEM persistence?

Overall, the results supported the literature showing that students from underrepresented populations were less likely to pursue and attain a degree in STEM (Ashford et al., 2016; Diemer & Li, 2012). However, it should be noted that Hispanic students with high achievement in STEM had a persistence rate similar to that of well-represented populations (i.e., Asian, White). Additionally, among non-high-achieving groups, Hispanic students had the highest rate of persistence. Factors such as motivation, self-efficacy, and school belonging might play a key role in these students' persistence in STEM fields.

Previous studies have shown that students from economically disadvantaged backgrounds perceive more barriers to postsecondary education (Turner et al., 2019) and are more likely to leave institutions before graduation (Holmes et al., 2018). In this study, I examined the importance of gender, ethnicity, and socioeconomic status using the random forest feature importance technique. Even though they were not the most significant variables of the study, the results indicated that socioeconomic status was the most important demographic variable.

The results confirm previous results showing that students from racially, culturally, economically, and linguistically diverse populations are less likely to pursue a STEM field

(Ashford et al., 2016; Diemer & Li, 2012). However, the study results also demonstrate that traditionally underrepresented populations can also persist similarly in STEM once they enter into STEM fields. Furthermore, being high-achieving might be a protective factor for students' readiness to succeed and persist in STEM since high-achieving students from low socioeconomic backgrounds persist similarly in STEM. Therefore, schools and educators need to provide opportunities to underrepresented students to pursue STEM fields (e.g., recruiting students for AP/IB courses), which will potentially help develop their talents in STEM.

Research Question 5: Which machine learning techniques can be used to identify variables influencing the early postsecondary persistence of high-achieving and non-high-achieving students in terms of classification models?

In this study, random forest and artificial neural network methods were used in predicting persistence in STEM fields. The results revealed that the accuracy of random forest and artificial neural network results was high in that the models had over 80% performance levels (Cardona et al., 2020). Additionally, sensitivity, specificity, and Receiver Operating Characteristics (ROC) Curve results showed that the models performed well, demonstrating that they could classify persistent and non-persistent students. Thus, the results of this study are in line with previous studies showing that random forest and artificial neural network models could be used to predict student persistence (Cardona et al., 2020; Thammasiri et al., 2014). Moreover, it should be noted that, overall, artificial neural network results were more accurate than random forest results. This finding aligns with previous research showing that artificial neural networks perform with greater accuracy than other models (Hodges & Mohan, 2019).

Compared to previously studied models, the models with SMOTE dataset performed at a higher level, over 82%, in this study. However, although the results with the SMOTE dataset provided good results, it should be noted that the random forest model could not handle an

imbalanced dataset and favored persistent students in STEM, which is a common problem in machine learning, resulting in overfitting of a major class. However, even with an imbalanced dataset, artificial neural network sensitivity was 75% for high-achieving and 95% non-high-achieving students' persistence. These results support the idea that artificial neural networks usually outperform other methods (Mason et al., 2018). Overall, the results provide evidence that the selected variables could predict persistence of high-achieving and non-high-achieving students and support the effectiveness of machine learning in educational research (Adejo & Connolly, 2018; Delen, 2010; Dissanayake et al., 2016; Kondo et al., 2017; Pereira et al., 2017).

Additionally, a nationally representative cohort of students was used in this study, which extended machine learning research on STEM persistence by expanding use of high school data to predict college persistence in STEM. Results of the study revealed that random forest and artificial neural network results can be good resources for the prediction of student persistence in STEM majors in the United States, and educators can use this information to develop strategies to support at-risk students and prevent future dropouts from STEM majors in college.

Limitations

There are several limitations in this study. First, even though I initially planned to use a much larger dataset, after feature engineering, the number of high-achieving and non-high-achieving students became much smaller compared to the initial dataset size. In general, larger datasets yield better results (Cardona, 2020); using a bigger dataset might have reduced the final number of variables and also might have yielded more reliable results.

Second, I used a publicly available dataset in which some variables were restricted, limiting the study's analyses in many ways. For instance, if AP/IB courses could have been investigated separately, more information could be gained based on each program's courses.

Finally, student persistence is a complex problem (Chen et al., 2018). In this study, I provided only random forest feature importance technique results to investigate the most significant variables of the study, which could be a limitation. Other techniques such as SHapley Additive exPlanations (SHAP) by Lundberg and Lee (2017) might have provide additional insights regarding each variable and its effect on STEM persistence.

Suggestions for Future Research

This study provides many insights into factors affecting both high-achieving and non-high-achieving students' persistence in STEM education. Further research can be conducted using a larger dataset. To deal with imbalanced datasets in machine learning, different techniques can be used such as under-sampling and over-sampling (Longadge et al., 2013; Thammasiri et al., 2014). A similar study with a larger sample and fewer variables included in the model can reduce the risk of overfitting of the model. Also in future research, other nationally representative datasets can be used to study the variables used in the study, which could provide further insights regarding the roles of demographics, cognitive factors, and non-cognitive factors in the prediction of student persistence in STEM fields in college.

Conclusion

With the goal of better understanding the nature and dynamics of STEM persistence, I examined the predictors of early postsecondary STEM persistence of high-achieving and non-high-achieving students. I presented the results for predicting and understanding STEM persistence using a national dataset collected by NCES. Unlike most previous studies, this study used machine learning methods to examine the persistence of high-achieving and non-high-achieving students in STEM fields in the United States

In most machine learning studies in post-secondary education, data are collected from university students (Baranyi et al., 2020; Zahedi et al., 2020). In this study, I used high school level factors to predict student persistence in STEM in post-secondary education. Educators should focus on all the variables included in the models because the models used in this study performed well, supporting that high school factors can be used to make strong predictions about student persistence in STEM in college using machine learning models. These models can be used not only to help researchers understand the factors that affect student persistence but to guide students along the right path to success (Alkhasawneh & Hargraves, 2014). Furthermore, because students who are more likely to be non-persistent in STEM can be identified right after high school, strategies can be developed to prevent their future dropping out from STEM education.

In general, larger datasets generate more reliable results in machine learning (Cardona, 2020). As noted, initially, this study had a larger dataset; however, the size decreased substantially after using feature engineering to answer research questions. In future research, it is recommended to work with a larger dataset and confirm the results of the study.

More research using machine learning is needed to determine factors influencing student persistence. Ultimately, the machine learning models used in this study can be used to develop strategies to increase students' persistence in STEM, such as creating services and programs to support them. Also, these models can help identify at-risk students and support their learning to avoid negative educational outcomes (Kučak et al., 2018).

Also, machine learning applications in education are limited; researchers can use machine learning to solve educational problems such as grading students and predicting student performance and achievement.

REFERENCES

Achter, J. A., Lubinski, D., Benbow, C. P., & Eftekhari-Sanjani, H. (1999). Assessing vocational preferences among gifted adolescents adds incremental validity to abilities: A discriminant analysis of educational outcomes over a 10-year interval. *Journal of Educational Psychology*, *91*(4), 777. https://doi.org/10.1037/0022-0663.91.4.777

Adelman, C. (1998). *Women and men of the engineering path: A model for analyses of undergraduate careers*. Office of Educational Research and Improvement, US Department of Education. https://files.eric.ed.gov/fulltext/ED419696.pdf

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

Alpaydin, E. (2004). *Introduction to machine learning*. MIT Press.

Alkhasawneh, R., & Hargraves, R. H. (2014). Developing a hybrid model to predict student first year retention in STEM disciplines using machine learning techniques. *Journal of STEM Education: Innovations and Research, 15*(3), 35-42.

Amit, Y. & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation, 9*(7), 1545-1588. https://doi.org/10.1162/neco.1997.9.7.1545

Anderson, K. A. (2016). Examining organizational practices that predict persistence among high-achieving black males in high school. *Teachers College Record, 118*(6), 1-26.

Aryee, M. (2017). *College students' persistence and degree completion in science, technology, engineering, and mathematics (STEM): The role of non-cognitive attributes of self-efficacy, outcome expectations, and interest.* (Publication No. 2246) [Doctoral Dissertation, Seton Hall University]. https://scholarship.shu.edu/dissertations/2246

Ashford, S. N., Lanehart, R. E., Kersaint, G. K., Lee, R. S., & Kromrey, J. D. (2016). STEM pathways: Examining persistence in rigorous math and science course taking. *Journal of Science Education and Technology*, *25*(6), 961-975. https://doi.org/10.1007/s10956-016-9654-0

Assouline, S. G., Ihrig, L. M., & Mahatmya, D. (2017). Closing the excellence gap: Investigation of an expanded talent search model for student selection into an extracurricular STEM program in rural middle schools. *Gifted Child Quarterly*, *61*(3), 250-261. https://doi.org/10.1177/0016986217701833

Aulck, L., Aras, R., Li, L., L'Heureux, C., Lu, P., & West, J. (2017). *S*TEM-ming the Tide: Predicting STEM attrition using student transcript data. *arXiv preprint arXiv:1708.09344*. https://arxiv.org/pdf/1708.09344.pdf

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*. https://arxiv.org/pdf/1606.06364.pdf

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Baranyi, M., Nagy, M., & Molontay, R. (2020, October 7-9). *Interpretable deep learning for university dropout prediction* [Conference Paper]. 21st Annual Conference on Information Technology Education, Virtual Event, USA. https://doi.org/10.1145/3368308.3415382

Boston, J. S., & Cimpian, A. (2018). How do we encourage gifted girls to pursue and succeed in science and engineering? *Gifted Child Today*, *41*(4), 196-207. https://doi.org/10.1177/1076217518786955

Camp, A. G., Gilleland, D., Pearson, C., & Putten, J. V. (2009). Women's path into science and engineering majors: a structural equation model. *Educational Research and Evaluation*, *15*(1), 63-77. https://doi.org/10.1080/13803610802591725

Cardona, T. A. (2020). *Development of a system architecture for the prediction of student success using machine learning techniques.* (Order No. 1198498983) [Doctoral dissertation, Missouri University of Science and Technology]. https://scholarsmine.mst.edu/doctoral_dissertations/2908

Cardona, T., Cudney, E. A., Hoerl, R., & Snyder, J. (2020). Data mining and machine learning retention models in higher education. *Journal of College Student Retention: Research, Theory & Practice*, *0*(0) 1–25. https://doi.org/10.1177/1521025120964920

Chen, X. (2009). *Students who study science, technology, engineering, and mathematics (STEM) in postsecondary education* (NCES 2009-161). National Center for Education Statistics. https://files.eric.ed.gov/fulltext/ED506035.pdf

Chen, X., & Soldner, M. (2013). *College students' paths into and out of STEM fields* (NCES 2014-001). National Center for Education Statistics. https://nces.ed.gov/pubs2014/2014001rev.pdf

Chen, Y., Johri, A., & Rangwala, H. (2018, March 7-9). *Running out of stem: A comparative study across stem majors of college students at-risk of dropping out early* [Conference Paper]. 8th International Conference on Learning Analytics and Knowledge, Sydney, NSW, Australia. https://doi.org/10.1145/3170358.3170410

Chimka, J. R., Reed-Rhoads, T., & Barker, K. (2007). Proportional hazards models of graduation. *Journal of College Student Retention: Research, Theory & Practice*, *9*(2), 221-232. https://doi.org/10.2190/CS.9.2.f

Chollet, F., & others. (2015). *Keras*. GitHub. https://github.com/fchollet/keras

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016, June 26-July 01). *Data cleaning: Overview and emerging challenges* [Conference Paper]. 2016 International Conference on Management of Data, San Francisco, CA, USA. https://doi.org/10.1145/2882903.2912574

College Board. (n.d.). *About AP*. https://aphighered.collegeboard.org/about-ap

College Board. (2011). *SAT percentile ranks for males, females, and total group.* https://secure-media.collegeboard.org/digitalServices/pdf/SAT-Mathemathics_Percentile_Ranks_2011.pdf

College Board. (2020). *AP cohort data report.* https://reports.collegeboard.org/pdf/2020-ap-cohort-data-report.pdf

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49(4),* 498-506. https://doi.org/10.1016/j.dss.2010.06.003

Diemer, M. A., & Li, C. H. (2012). Longitudinal roles of precollege contexts in low-income youths' postsecondary persistence. *Developmental Psychology*, *48*(6), 1686–1693. https://doi.org/10.1037/a0025347

Dissanayake, H., Robinson, D., & Al-Azzam, O. (2016). Predictive Modeling for Student Retention at St. Cloud State University. *Proceedings of the International Conference on Data Mining (DMIN)* (pp. 215-221). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp). http://worldcomp-proceedings.com/proc/p2016/DMI8024.pdf

Duprey, M.A., Pratt, D.J., Jewell, D.M., Cominole, M.B., Fritch, L.B., Ritchie, E.A., Rogers, J.E., Wescott, J.D., Wilson, D.H. (2018). *High school longitudinal study of 2009 (HSLS:09): Base-year to second follow-up data file documentation* (NCES 2018-140). National Center for Education Statistics. https://nces.ed.gov/pubs2018/2018140.pdf

Dweck, C. S. (2002). The development of ability conceptions. In Wigfield, A., & Eccles, J. S. (Eds.), *Development of achievement motivation* (pp. 57-88). Academic Press. https://doi.org/10.1016/B978-012750053-9/50005-X

Dweck, C. (2007). Is math a gift? Beliefs that put females at risk. In S. J. Ceci & W. M. Williams (Eds.), *Why aren't more women in science?* (pp. 47-55). American Psychological Association. https://doi.org/10.1037/11546-004

Eccles, J. S., Vida, M. N., & Barber, B. (2004). The relation of early adolescents' college plans and both academic ability and task-value beliefs to subsequent college enrollment. *The Journal of Early Adolescence, 24*(1), 63–77. https://doi.org/10.1177/0272431603260919

Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information Sciences, 505*, 32-64. https://doi.org/10.1016/j.ins.2019.07.070

Eris, O., Chachra, D., Chen, H. L., Sheppard, S., Ludlow, L., Rosca, C., Bailey, T., & Toye, G. (2010). Outcomes of a longitudinal administration of the persistence in engineering survey. *Journal of Engineering Education*, *99*(4), 371-395. https://doi.org/10.1002/j.2168-9830.2010.tb01069.x

French, B. F., Immekus, J. C., & Oakes, W. C. (2005). An examination of indicators of engineering students' success and persistence. *Journal of Engineering Education*, *94*(4), 419-425. https://doi.org/10.1002/j.2168-9830.2005.tb00869.x

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and tensorflow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Green, A., & Sanderson, D. (2018). The roots of STEM achievement: An analysis of persistence and attainment in STEM majors. *The American Economist*, *63*(1), 79-93. https://doi.org/10.1177/0569434517721770

Hastie, T., Tibshirani, T., & Friedman, J. (2009). *The elements of statistical learning*. Springer.

Heilbronner, N. N. (2011). Stepping onto the STEM pathway: Factors affecting talented students' declaration of STEM majors in college. *Journal for the Education of the Gifted*, *34*(6), 876-899. https://doi.org/10.1177/0162353211425100

Heilbronner, N. N. (2013). The STEM pathway for women: What has changed? *Gifted Child Quarterly*, *57*(1), 39-55. https://doi.org/10.1177/0016986212460085

Heydt, M. (2017). *Learning pandas*. Packt Publishing Ltd.

Higdem, J. L., Kostal, J. W., Kuncel, N. R., Sackett, P. R., Shen, W., Beatty, A. S., & Kiger, T. B. (2016). The role of socioeconomic status in SAT–freshman grade relationships across gender and racial subgroups. *Educational Measurement: Issues and Practice*, *35*(1), 21-28. https://doi.org/10.1111/emip.12103

Ho, T. K. (1995, August 14-16). *Random decision forests* [Conference Paper]. 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada. https://doi.org/10.1109/ICDAR.1995.598994

Hodges, J., & Mohan, S. (2019). Machine learning in gifted education: A demonstration using neural networks. *Gifted Child Quarterly*, *63*(4), 243-252. https://doi.org/10.1177/0016986219867483

Holmes, K., Gore, J., Smith, M., & Lloyd, A. (2018). An integrated analysis of school students' aspirations for STEM careers: Which student and school factors are most predictive? *International Journal of Science and Mathematics Education*, *16*(4), 655-675. https://doi.org/10.1007/s10763-016-9793-z

Ingels, S.J., Pratt, D.J., Herget, D.R., Burns, L.J., Dever, J.A., Ottem, R., Rogers, J.E., Jin, Y., and Leinwand, S. (2011). *High school longitudinal study of 2009 (HSLS:09): Base-year data file documentation* (NCES 2011-328). National Center for Education Statistics. https://nces.ed.gov/pubs2014/2014361.pdf

Ingels, S.J., Pratt, D.J. Herget, D.R., Dever, J.A., Fritch, L.B., Ottem, R., Rogers, J.E., Kitmitto, S., and Leinwand, S. (2013). *High school longitudinal study of 2009 (HSLS:09): Base year to first follow-up data file documentation* (NCES 2014- 361). National Center for Education Statistics. https://files.eric.ed.gov/fulltext/ED565693.pdf

Ingels, S.J., Pratt, D.J., Herget, D., Bryan, M., Fritch, L.B., Ottem, R., Rogers, J.E., and Wilson, D. (2015). *High school longitudinal study of 2009 (HSLS:09) 2013 update and high school transcript data file documentation* (NCES 2015-036). National Center for Education Statistics. https://nces.ed.gov/pubs2015/2015036.pdf

International Baccalaureate Organization (2017). *The history of the IB.* https://www.ibo.org/globalassets/digital-toolkit/presentations/1711-presentation-history-of-the-ib-en.pdf

Kerr, B. A., Multon, K. D., Syme, M. L., Fry, N. M., Owens, R., Hammond, M., & Robinson-Kurpius, S. (2012). Development of the distance from privilege measures: A tool for understanding the persistence of talented women in STEM. *Journal of Psychoeducational Assessment*, *30*(1), 88-102. https://doi.org/10.1177/0734282911428198

Kim, A. Y., Sinatra, G. M., & Seyranian, V. (2018). Developing a STEM identity among young women: a social identity perspective. *Review of Educational Research*, *88*(4), 589-625. https://doi.org/10.3102/0034654318779957

Kondo, N., Okubo, M., & Hatanaka, T. (2017, July 9-13). *Early detection of at-risk students using machine learning based on LMS log data* [Conference Paper]. 6th IIAI International Congress on Advanced Applied Informatics, Hamamatsu, Japan. https://doi.org/10.1109/IIAI-AAI.2017.51

Kučak, D., Juričić, V., & Đambić, G. (2018). Machine learning in education-A survey of current research trends. In B. Katalinic (Ed.), *Proceedings of the 29th DAAAM International Symposium* (pp. 0406-0410). https://doi.org/10.2507/29th.daaam.proceedings.059

Lewis, K. L., Stout, J. G., Finkelstein, N. D., Pollock, S. J., Miyake, A., Cohen, G. L., & Ito, T. A. (2017). Fitting in to move forward: Belonging, gender, and persistence in the physical sciences, technology, engineering, and mathematics (pSTEM). *Psychology of Women Quarterly*, *41*(4), 420-436. https://doi.org/10.1177/0361684317720186

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior*, *45*(1), 79-122. https://doi.org/10.1006/jvbe.1994.1027

Lent, R. W., Brown, S. D., & Hackett, G. (2000). Contextual supports and barriers to career choice: A social cognitive analysis. *Journal of Counseling Psychology*, *47*(1), 36. https://doi.org/10.1037/0022-0167.47.1.36

Longadge, R., & Dongre, S. (2013). *Class imbalance problem in data mining review*. https://arxiv.org/pdf/1305.1707.pdf

Lubinski, D., Webb, R. M., Morelock, M. J., & Benbow, C. P. (2001). Top 1 in 10,000: A 10-year follow-up of the profoundly gifted. *Journal of Applied Psychology*, *86*(4), 718. http://doi.org/10.1037//0021-9010.86.4.718

Lundberg, S., & Lee, S. I. (2017). *A unified approach to interpreting model predictions*. https://arxiv.org/pdf/1705.07874.pdf

Maltese, A. V., Melki, C. S., & Wiebke, H. L. (2014). The nature of experiences responsible for the generation and maintenance of interest in STEM. *Science Education*, *98*(6), 937-962. https://doi.org/10.1002/sce.21132

Maltese, A. V., & Tai, R. H. (2011). Pipeline persistence: Examining the association of educational experiences with earned degrees in STEM among US students. *Science Education*, *95*(5), 877-907. https://doi.org/10.1002/sce.20441

Mason, C., Twomey, J., Wright, D., & Whitman, L. (2018). Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a backpropagation neural network and logistic regression. *Research in Higher Education*, *59*(3), 382-400. https://doi.org/10.1007/s11162-017-9473-z

Maseleno, A., Sabani, N., Huda, M., Ahmad, R., Jasmi, K. A., & Basiron, B. (2018). Demystifying learning analytics in personalised learning. *International Journal of Engineering & Technology*, *7*(3), 1124-1129. https://doi.org/10.14419/ijet.v7i3.9789

Mattern, K. D., Marini, J. P., & Shaw, E. J. (2013). *Are AP students more likely to graduate from college on time?* College Board. https://files.eric.ed.gov/fulltext/ED556464.pdf

Mendez, G., Buskirk, T. D., Lohr, S., & Haag, S. (2008). Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests. *Journal of Engineering Education*, *97*(1), 57-70. https://doi.org/10.1002/j.2168-9830.2008.tb00954.x

Midi, H., Sarkar, S. K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, *13*(3), 253-267. https://doi.org/10.1080/09720502.2010.10700699

Min, Y., Zhang, G., Long, R. A., Anderson, T. J., & Ohland, M. W. (2011). Nonparametric survival analysis of the loss rate of undergraduate engineering students. *Journal of Engineering Education*, *100*(2), 349-373. https://doi.org/10.1002/j.2168-9830.2011.tb00017.x

Müller, A. C., Guido, S. (2016). *introduction to machine learning with python: A guide for data scientists.* O'Reilly Media.

National Center for Education Statistics (2017). *The condition of education.* https://nces.ed.gov/programs/coe/pdf/coe_tsc.pdf

National Science Board. (2018). *Science and engineering indicators 2018.* https://www.nsf.gov/statistics/2018/nsb20181/assets/nsb20181.pdf

Nicholls, G. M., Wolfe, H., Besterfield-Sacre, M., & Shuman, L. J. (2010). Predicting STEM degree outcomes based on eighth grade data and standard test scores. *Journal of Engineering Education*, *99*(3), 209-223. https://doi.org/10.1002/j.2168-9830.2010.tb01057.x

Nicholls, G. M., Wolfe, H., Besterfield-Sacre, M., Shuman, L. J., & Larpkiattaworn, S. (2007). A method for identifying variables for predicting STEM enrollment. *Journal of Engineering Education*, *96*(1), 33-44. https://doi.org/10.1002/j.2168-9830.2007.tb00913.x

Nugent, G., Barker, B., Welch, G., Grandgenett, N., Wu, C., & Nelson, C. (2015). A model of factors contributing to STEM learning and career orientation. *International Journal of Science Education, 37*(7), 1067-1088. https://doi.org/10.1080/09500693.2015.1017863

Okubo, F., Shimada, A., Yamashita, T., & Ogata, H. (2017, March 13-17). *A neural network approach for students' performance prediction* [Conference Poster]. 7th International Learning Analytics and Knowledge Conference, Vancouver, BC, Canada. https://doi.org/10.1145/3027385.3029479

Olszewski-Kubilius, P. (2006). The role of summer enrichment programs in developing the talents of gifted students. In VanTassel-Baska (Ed.), *Serving gifted learners beyond the traditional classroom* (pp. 13-35). Prufrock Press.

Pavleković, M., Zekić-Sušac, M., & Đurđević, I. (2011). A neural network model for predicting children's mathematical gift. *Croatian Journal of Education, 13*(1), 10-41.

Pereira, R. T., & Zambrano, J. C. (2017, December 18-21). Application of decision trees for detection of student dropout profiles [Conference Paper]. *16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico.* https://doi.org/10.1109/ICMLA.2017.0-107

Pilchen, A., Caspary, K., & Woodworth, K. (2019). *Postsecondary outcomes of IB diploma programme graduates in the US*. https://ibo.org/globalassets/publications/ib-research/dp/us-postsecondary-outcomes-final-report.pdf

Richert, W., Coelho, L. P. (2013). *Building machine learning systems with Python*. Packt Publishing.

Rosebrock, A. (2019). *Deep learning for computer vision with Python: Starter bundle* (3rd ed.). PyImageSearch.

Saeys, Y., Abeel T., Van de Peer Y. (2008) Robust Feature Selection Using Ensemble Feature Selection Techniques. In: Daelemans W., Goethals B., Morik K. (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 313-325). Springer. https://doi.org/10.1007/978-3-540-87481-2_21

Sage, A. J., Cervato, C., Genschel, U., & Ogilvie, C. A. (2018). Combining academics and social engagement: A major-specific early alert method to counter student attrition in science, technology, engineering, and mathematics. *Journal of College Student Retention: Research, Theory & Practice*, *22*(4), 611-626. https://doi.org/10.1177/1521025118780502

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210-229. https://doi.org/10.1147/rd.33.0210

Shaw, E. J., & Barbuti, S. (2010). Patterns of persistence in intended college major with a focus on STEM majors. *NACADA Journal*, *30*(2), 19-34. https://doi.org/10.12930/0271-9517-30.2.19

Smith, K., Jagesic, S., Wyatt, J., & Ewing, M. (2018). *AP STEM participation and postsecondary STEM outcomes: Focus on underrepresented minority, first-generation, and female students.* College Board. https://files.eric.ed.gov/fulltext/ED581514.pdf

Simon, R. A., Aulls, M. W., Dedic, H., Hubbard, K., & Hall, N. C. (2015). Exploring student persistence in STEM programs: A motivational model. *Canadian Journal of Education*, *38*(1), 1-27.

Steenbergen-Hu, S., & Olszewski-Kubilius, P. (2017). Factors that contributed to gifted students' success on stem pathways: The role of race, personal interests, and aspects of high school experience. *Journal for the Education of the Gifted*, *40*(2), 99-134. https://doi.org/10.1177/0162353217701022

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, *41*(2), 321-330. https://doi.org/10.1016/j.eswa.2013.07.046

Turner, S. L., Joeng, J. R., Sims, M. D., Dade, S. N., & Reid, M. F. (2019). SES, gender, and STEM career interests, goals, and actions: A test of SCCT. *Journal of Career Assessment*, *27*(1), 134-150. https://doi.org/10.1177/1069072717748665

Tyson, W., Lee, R., Borman, K. M., & Hanson, M. A. (2007). Science, technology, engineering, and mathematics (STEM) pathways: High school science and math coursework and postsecondary degree attainment. *Journal of Education for Students Placed at Risk*, *12*(3), 243-270. https://doi.org/10.1080/10824660701601266

U.S Census Bureau (n.d.). *QuickFacts.* https://www.census.gov/quickfacts/fact/table/US/POP010210

U.S. Department of Commerce. (2011). *Women in STEM: A gender gap to innovation*. http://www.esa.doc.gov/sites/default/files/reports/documents/womeninstemagaptoinnovation8311.Pdf

Wai, J., Lubinski, D., Benbow, C. P., & Steiger, J. H. (2010). Accomplishment in science, technology, engineering, and mathematics (STEM) and its relation to STEM educational dose: A 25-year longitudinal study. *Journal of Educational Psychology*, *102*(4), 860. https://doi.org/10.1037/a0019454

Wang, M. T., & Degol, J. L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review*, *29*(1), 119-140. https://doi.org/10.1007/s10648-015-9355-x

Wang, X. (2013). Why students choose STEM majors: Motivation, high school learning, and postsecondary context of support. *American Educational Research Journal*, *50*(5), 1081-1121. https://doi.org/10.3102/0002831213488622

Watkins, J., & Mazur, E. (2013). Retaining students in science, technology, engineering, and mathematics (STEM) majors. *Journal of College Science Teaching*, *42*(5), 36-41.

Yi, S. (2018). *Postsecondary stem paths of high-achieving students in math and science: A longitudinal multilevel investigation of their selection and persistence* (Publication No. 10829155) [Doctoral Dissertation, Purdue University]. ProQuest Dissertations and Theses Global.

Zahedi, L., Lunn, S. J., Pouyanfar, S., Ross, M. S., & Ohland, M. W. (2020, June 22-26). *Leveraging machine learning techniques to analyze computing persistence in undergraduate programs* [Conference Paper]. 2020 ASEE Virtual Annual Conference, Virtual Event, USA. https://doi.org/10.18260/1-2--34921

Zhang, S. (2012). Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software*, *85*(11), 2541-2552. http://dx.doi.org/10.1016/j.jss.2012.05.073

Zhang, G., Anderson, T. J., Ohland, M. W., & Thorndyke, B. R. (2004). Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study. *Journal of Engineering Education*, *93*(4), 313-320. https://doi.org/10.1002/j.2168-9830.2004.tb00820.x