Applications of Machine Learning Algorithms for Examining the Impact of COVID-19 on the Dropout Rate for High Schools in the State of Illinois.

Dissertation Manuscript

Submitted to National University

School of Technology

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

by

CLAUDE MARCELIN NGANTCHOU

San Diego, California

December 2025

Abstract

High school dropout remains a persistent and pressing issue in the United States and globally. This quantitative, non-experimental study aimed to develop a predictive model for high school dropout rates in the state of Illinois before and during the COVID-19 pandemic, covering the years 2017–2019 and 2020–2022, respectively. Publicly available datasets from the Illinois State Board of Education were analyzed using multiple linear regression, random forest, and XGBoost models to assess the impact of the pandemic and to identify which school-level features most strongly predict dropout outcomes. The study applied the CRISP-DM framework and interpreted results through the lens of survival analysis theory to address the problem of academic attrition over time. Research questions and hypotheses were tested using the three predictive models. The analysis identified mobility rate, COVID-19 period, *and* low-income enrollment as the most influential predictors of high school dropout, with mobility rate emerging as the top signal across models. Model performance was evaluated using R^2, mean absolute error (*MAE*), and root mean squared error (*RMSE*). The XGBoost model offered the best balance of predictive accuracy and computational efficiency, making it the most effective and preferred model for this study. Recommendations for future research are grounded in the study's predictive scope and methodological limitations. Proposed next steps include evaluating model performance over extended timeframes, incorporating post–COVID-19 data, and exploring additional demographic and school-level predictors using time-aware validation and stratified replication. These extensions will strengthen the generalizability and practical value of predictive modeling for dropout prevention in educational settings.

Acknowledgements

First and foremost, I acknowledge and thank my dissertation Chair, Dr. Irene Tsapara. Without Dr. Tsapara's continuous encouragement, guidance, and feedback, it would have been very difficult for me to complete this dissertation manuscript. The support of my subject matter expert, Dr. Aeron Zentner, was very valuable to the achievement of this milestone. I also acknowledge the support of my academic reader, Dr. Javier Rodriguez. I want to acknowledge my wife, Lydie Ngantchou. Her understanding, patience, and commitment to taking care of our little children Marie, Mathieu, and Joseph as I stayed up late at night to work on my dissertation could not go unnoticed. Finishing this manuscript without her help would not have been possible. I also thank our oldest child Kassandra and my uncle Joseph for their encouragement. Finally, I will thank my coworkers for the crucial and sustained role they played in providing me with motivation during this doctoral journey.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Education is a powerful force that drives economic growth in the United States, but many students dropped out of high school, making it difficult for the government to manage the consequences. Dropouts were more likely to be unemployed, rely on welfare, pay less in taxes, and receive tax-funded benefits (Foreman-Murray et al., 2022; Imola & Krisztina, 2019; Reeves, 2021). In 2018, half a million students dropped out of high school in the United States (Institute of Education Sciences, 2021). Miller (2011) stated that the number of students who left high school in the United States without completing their education surpassed 1.2 million annually. Lee and Polachek (2018) stated that the United States government was striving to address the issue of dropout, as it would improve its citizens' living conditions.

While school dropout rates were a concern in many countries, the coronavirus disease (COVID-19) pandemic exacerbated this issue (Shuja et al., 2022). Numerous studies examined the impacts of COVID-19 on educational systems in different countries (Colpo et al., 2024; Lichand et al., 2022; Shuja et al., 2022). The findings of these studies showed that educational institutions faced staff shortages, closures, and high rates of absenteeism and school dropout. The COVID-19 pandemic disrupted the learning process to a great extent. This interruption in traditional education resulted in significant student learning losses and adversely impacted the economic stability of vulnerable families. The ensuing financial difficulties led to a sharp increase in the school dropout rate (Rekha et al., 2023). A recent study conducted by the United Nations Educational, Scientific and Cultural Organization (UNESCO) showed that the COVID-19 pandemic caused 24 million learners to consider dropping out of school (Song et al., 2023).

Various countries capitalized on education to improve the well-being of their societies, as it allows citizens to acquire knowledge and skills that are key determinants of economic growth and welfare (Ansong et al., 2023). Chitsamatanga and Rembe (2020) stated that education was a means many countries used to catapult their development. Nations with high rates of education completion have less crime, better overall health, and higher levels of civic involvement (Gausel & Bourguignon, 2020). Nygren et al. (2020) emphasized education's vital role in promoting global goals. Results from various studies show that education plays a crucial role in establishing peaceful and healthy relationships among communities (Hasani & Kamberi, 2024; Mtey, 2024).

Gausel and Bourguignon (2020) stated that many communities considered individuals who dropped out of high school as people who had failed in life. Learners with disabilities represented a significant portion of the student population (Papadakaki et al., 2022). Students with disabilities tended to drop out of high school at higher rates than those without disabilities (McCauley, 2017). Irwin et al. (2021) stated that the number of students with disabilities who dropped out of high school was more than twice the number of students without disabilities. Approximately 11.7% of students with disabilities left high school each year (Hussar et al., 2020). The overall high school dropout rate in the U.S. in 2021 was 5.2% (National Center for Education Statistics, 2023).

The U.S. government was concerned about the high dropout rate, which negatively impacted the lives of its citizens (Hussar et al., 2020; McFarland et al., 2018). This high dropout rate decreased literacy levels and increased unemployment (Reeves, 2021). The United States government provided funds for the education of its young citizens to ensure their schooling (U.S. Department of Education, 2022). This persistent issue of high school dropouts affected

areas beyond education (Imola & Krisztina, 2019). Consequently, it was important to address this issue. This study measured the predictability of the high school dropout rate based on factors that included student mobility rate, student attendance rate, the rate of students with Individualized Education Plans (IEPs), and the rate of students from low-income families. The study also examined the extent to which a significant difference existed in high school dropout rates before and during the COVID-19 pandemic in the state of Illinois.

Statement of the Problem

The problem addressed in this study was the high rate of dropout among high school students in Illinois from 2017 to 2022. Munk et al. (2021) stated that dropout rates for both regular and disabled students were gaining attention in the United States. Students who left high school before graduating were more likely to be involved in crime and incarceration (Gausel & Bourguignon, 2020). These students often experienced less favorable outcomes in many sectors of life (Gausel & Bourguignon, 2020; McCauley, 2017). The United States held one of the highest high school dropout rates in the world (Ressa & Andrews, 2022).

High school dropout rates had a negative impact on social stability and economic development (Hussar et al., 2020; McFarland et al., 2018). Students who chose to drop out were at increased risk of unemployment. Lee and Polachek (2018) found that students who dropped out of school were more likely to receive government assistance. Additionally, students who dropped out earned $183 less per week than those with a high school diploma (U.S. Bureau of Labor Statistics, 2021).

Many studies, including McFarland et al. (2018) and Irwin et al. (2021), examined high school dropout rates with limited focus on students' decision-making processes. Education stakeholders, including teachers, administrators, staff members, students, parents, community

members, and elected officials, were searching for solutions (McFarland et al., 2020). Wan (2022) identified a gap in the literature relating to student motivation to drop out. Foreman-Murray et al. (2022) reinforced the need for further research by highlighting the relationship between school engagement and student dropout. Despite numerous studies exploring the causes of dropout, the issue persisted. If not addressed, the economic and social consequences would likely intensify (Lee & Polachek, 2018). Therefore, the problem required thorough investigation to provide school officials with insights that might help reduce the dropout rate significantly.

Purpose of the Study

The purpose of this quantitative study was to determine a predictive model for high school dropout rates before and during COVID-19 in the state of Illinois from 2017–2019 and 2020–2022, respectively. Predictor variables included student demographics, geolocation, and other internal school factors. The high school dropout rate served as the response variable. As little was known about the factors contributing to high dropout rates, this study explored the predictive relationship between these variables and dropout rates. Designing and demonstrating this predictive model helped address the research questions and provided potential solutions to mitigate the persistent problem of high school dropout.

Introduction to Theoretical Framework

The theoretical framework for this study was based on a combination of survival analysis and the Cross-Industry Standard Process for Data Mining (CRISP-DM). Survival analysis originated in the 17th century when John Graunt (1662) produced the first life table, which is still used to answer a wide range of research questions involving time or duration. Survival analysis is widely used in various industries due to its advantages in processing censored data

(Peset et al., 2020). Researchers use survival analysis to study the duration from the start time of a participant's involvement until the occurrence of an event, the conclusion of the study, or participant dropout (Boualaphet & Goto, 2020). It is used to indicate whether an individual has experienced an event of interest by a specific time. In this study, survival analysis theory was applied to examine the dropout rate at the high school level, specifically, whether students "survived" academically over a given time span.

CRISP-DM is a standardized framework that helps researchers determine how to approach a data-related problem. It was developed in the 1990s as a methodology for data mining projects across industries. CRISP-DM remains the de facto standard for data mining and data science projects (Martinez-Plumed et al., 2021). The process consists of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. In this study, the CRISP-DM process was used to predict the high school dropout rate.

This study contributed to existing research, including McFarland et al. (2018) and Irwin et al. (2021), on high school dropout by applying survival analysis to the attitudes and learning behaviors of students who dropped out, which, in turn, influenced the school's dropout rate. According to Jin (2023), positive attitudes and good learning behaviors support students' autonomy, relevance, and relatedness to the material, contributing to their academic success. The inclusion of predictor variables such as student demographics, geolocation, and other internal school factors yielded highly accurate results, as these variables reflect the underlying attitudes and learning behaviors of students (Alspaugh, 2000; Cocoradă et al., 2021; de la Cruz Orozco & Heredia Rubio, 2019; Roderick et al., 2012; South et al., 2007). These variables provided evidence of student motivation, which correlates strongly with classroom engagement (Girgin & Cabaroğlu, 2021). Motivation among students who dropped out tended to decline over time.

However, administrators, instructors, field experts, and support staff were not always aware of the reasons students chose to drop out (Yılmaz & Karataş, 2022). Low levels of school participation were associated with an increased likelihood of dropout.

Introduction to Research Methodology and Design

A quantitative, non-experimental predictive modeling method was used to conduct this study. Predictive modeling was appropriate, as the research question addressed the predictive relationship between multiple variables. The study aimed to determine a predictive model for high school dropout rates based on student demographic data, geolocation, and internal school factors.

Students often choose actions based on the perceived desirability of outcomes. To explore this, the study analyzed existing secondary data collected by the Illinois State Board of Education (ISBE) from 2017 to 2022. This data pertained to high schools and school districts and was used to examine the issue of dropout.

A representative sample size was used in the analysis. Ensuring that statistical assumptions were met was essential. The goal of this study was to apply a predictive model for the high school dropout rate based on demographic, geolocation, and internal school factors. The research question aligned with both the theoretical framework and the predictive modeling method employed.

Research Questions

RQ1

To what extent, if any, the demographic, geolocation and other internal factors of the schools predict the high school dropout rate in the state of Illinois?

RQ2

Which predictor variables significantly influence the prediction of the high school dropout rate in the state of Illinois?

Hypotheses

$H1_0$:

There is no statistically significant predictive relationship between the demographic, geolocation, and other internal factors of the schools and the high school dropout rate in the state of Illinois.

H1a:

There is a statistically significant predictive relationship between the demographic, geolocation, and other internal factors of the schools and the high school dropout rate in the state of Illinois

$H2_0$:

There are no predictor variables that significantly influence the prediction of the high school dropout rate in the state of Illinois.

H2a:

There are predictor variables that significantly influence the prediction of the high school dropout rate in the state of Illinois.

Significance of the Study

Numerous studies examined dropout among both regular students and those with disabilities (Alspaugh, 2000; Hussar et al., 2020; Irwin et al., 2021; McFarland et al., 2018). However, limited research was conducted on factors that led students to consider dropping out based on demographic, geolocation, and other internal school variables. Therefore, this study

examined the high school dropout rate based on those predictors in the state of Illinois from 2017–2022. This research provided insights that could be used to explore ways to significantly reduce the dropout rate for both regular students and those with disabilities.

The Individuals with Disabilities Education Act (IDEA) is a United States law passed in 1990 to replace the Education for All Handicapped Children Act. It was amended in December 2004 to ensure equal access to education. IDEA ensures that every child with a disability across the country has access to special education services that meet their state's learning standards. School dropout has remained a major concern for the United States (Lee & Polachek, 2018). Students with disabilities tend to disengage more quickly than their non-disabled peers (McCauley, 2017). This study demonstrated that predicting dropout for both regular and disabled students was essential, given their active participation in the learning process. The study focused on predicting the high school dropout rate based on demographic, geolocation, and internal school factors in Illinois from 2017–2022. However, few studies have investigated predictive dropout modeling based on these variables in Illinois during this timeframe.

Definition of Key Terms

Disability

Students who participate in specialized educational programs due to physical, emotional, or mental conditions in order to meet state learning standards (Lequia et al., 2023).

English language learners (ELLs)

Students identified through a screening process as eligible for bilingual education or English as a second language services who have not yet achieved proficiency in English (Illinois State Board of Education [ISBE], 2023).

Individuals with Disabilities Education Act

A federal law implemented across all states to guarantee access to public education for students with disabilities (Voulgarides & Barrio, 2021).

Motivation

The capacity to demonstrate goal-oriented behaviors and complete tasks (André et al., 2023).

Prediction

A statistical technique using machine learning to analyze historical and current data in order to forecast future outcomes (Meng et al., 2023).

Rate of students from low-income families

The percentage of students from families with limited financial resources (Rose et al., 2012).

Rate of students with IEP

The proportion of students receiving specialized instruction and services through an Individualized Education Plan (Churchill et al., 2021).

School district size

The geographic area governed by a school board that includes all public primary and secondary schools within that boundary (Haeberlein et al., 2021).

Special education

Educational programs funded by the U.S. government designed to support students with disabilities at no cost to their families (Lequia et al., 2023).

Student attendance rate

The proportion of students attending school during regularly scheduled hours (Tomaszewski et al., 2023).

Student mobility rate

The rate at which students transfer in and out of a school during a given academic period (LeBoeuf & Fantuzzo, 2018).

Summary

This quantitative study examined the predictive relationships among several variables. Its purpose was to determine a predictive model for the high school dropout rate based on demographic, geolocation, and internal school factors. A predictive modeling approach was chosen because it enables analysis of past data to forecast future outcomes. This model could also be retrained with new data to refine predictions over time (Orlova, 2021).

Machine learning models, including linear regression, random forest, and extreme gradient boosting (XGBoost), were used to predict high school dropout based on demographic, geolocation, and internal school factors (Adasme et al., 2023; Manoj et al., 2023). Linear regression was used to analyze the relationship between predictors and the dropout rate. This technique allowed the researcher to assess total model variance and the contribution of each predictor variable.

A random forest model was used to combine the predictions from multiple decision trees, improving prediction accuracy and reducing overfitting. To further strengthen predictive performance, the study incorporated a XGBoost regression model. XGBoost is a highly scalable, efficient machine learning algorithm that constructs a sequence of decision trees, each correcting errors from the previous iteration. It uses regularization techniques to prevent overfitting and is well suited for identifying complex relationships and feature importance. Including XGBoost

allowed for comparison and validation of results across a range of modeling techniques, improving the reliability and practical value of the findings.

Chapter 2: Literature Review

This quantitative exploratory study aimed to understand the implications of COVID-19 and the demographic, geolocation, and other internal school factors on high school dropout rates in the state of Illinois from 2020 to 2022. The researcher built a predictive model to examine the influence of these variables on dropout rates before COVID-19 (2017–2019) and during COVID-19 (2020–2022). The problem addressed in this study was the high dropout rate among high school students in Illinois (ISBE, 2023). Dropping out of high school has been a persistent and deeply concerning issue in the United States (Foreman-Murray et al., 2022; Lee & Polachek, 2018). High school dropout rates have far-reaching consequences, affecting not only individuals who leave school prematurely but also society as a whole.

This literature review provides an overview of the problem of high school dropout, highlighting the need for targeted interventions and predictive models to address this issue effectively. High school dropout has been widely recognized as a societal challenge with severe consequences (Imola & Krisztina, 2019). In 2018, approximately 55% of high school dropouts were unemployed (U.S. Bureau of Labor Statistics, 2021). These individuals face limited job opportunities and reduced earning potential (Lee & Polachek, 2018). The absence of a high school diploma is associated with higher chances of incarceration and subsequent difficulties reintegrating into society (Gausel & Bourguignon, 2020).

The United States has witnessed a staggering number of students leaving high school without completing their education, surpassing 1.2 million annually (Miller, 2011). In 2022, there were about 2.1 million students between the ages of 16 and 24 in the United States who had dropped out of high school (National Center for Education Statistics, 2024). The U.S. government has acknowledged the urgency of addressing this issue, given its direct impact on

the nation's economic and social well-being (Lee & Polachek, 2018). This study incorporated existing literature on high school dropout to enrich the analysis derived from the quantitative approach. Database searches were conducted using relevant keywords.

Education plays a pivotal role in the development of societies. Countries that prioritize education tend to experience lower crime rates, improved overall health, and increased civic engagement (Chitsamatanga & Rembe, 2020; Gausel & Bourguignon, 2020). Education is also closely linked to achieving global goals such as ending poverty and protecting the planet (Nygren et al., 2020). Promoting education leads to a better society, with well-informed citizens contributing to national development (Chitsamatanga & Rembe, 2020). Lower crime rates emerge as education empowers individuals and reduces the likelihood of engaging in illegal or antisocial behavior. Access to quality education equips people with critical thinking and problem-solving skills, which contribute to the overall safety and stability of society (Nygren et al., 2020). When education becomes a national priority, its positive effects reverberate across social domains.

In considering the relationship between education and overall health, including physical and mental health, it is essential to recognize the inclusive role of education, particularly for students with disabilities. Inclusive education not only imparts knowledge but also fosters healthier, more equitable societies (Papadakaki et al., 2022). For students with disabilities, education empowers informed health choices (Nygren et al., 2020; Papadakaki et al., 2022). Inclusive educational settings provide tailored content and resources to support healthy behaviors. Irwin et al. (2021) noted that this includes adapted lessons on physical activity, nutrition, and mental health designed to meet the needs of diverse learners.

By integrating health education within an inclusive academic framework, students with disabilities gain a better understanding of hygiene, nutrition, and lifestyle decisions suited to their needs (Kang & Chang, 2019). These benefits extend to families and communities (Papadakaki et al., 2022). For example, instruction on adaptive physical activities and proper nutrition has been shown to promote healthier habits across peer and social groups. Kang and Chang (2019) further demonstrated that inclusive education systems improve access to healthcare. Educated individuals, regardless of ability, are more likely to understand medical information, advocate for their needs, and access appropriate care. Thus, education ensures that students with disabilities can develop essential health literacy and communication skills, contributing to a proactive approach to wellness.

Conversely, learners who drop out of school face multiple disadvantages. Individuals who drop out are frequently stigmatized as failures (Gausel & Bourguignon, 2020). Among those who drop out, students with disabilities represent a disproportionately large share (Papadakaki et al., 2022). These students are more likely to exit school prematurely than their non-disabled peers (Irwin et al., 2021; McCauley, 2017). The dropout rate among students with disabilities is more than twice that of their peers without disabilities, a deeply troubling pattern.

Many researchers have investigated the issue of high school dropout among disabled and non-disabled students for years. However, the problem remains complex and multifaceted (Hussar et al., 2020). High school dropout among regular and disabled students affects a wide range of stakeholders, including teachers, administrators, students, parents, communities, and elected officials (McFarland et al., 2020). Researchers have examined various facets, including the educational structure, support systems, individual learning needs, societal perceptions, and systemic barriers, only to find a web of interconnected issues that contribute to the persisting

high dropout rates (Hussar et al., 2020). According to Hussar et al. (2020), this complexity arises from the interaction of diverse elements such as the lack of tailored educational support, inaccessible learning environments, societal attitudes, and policy gaps. All these factors intersect to create formidable hurdles for both regular and disabled students.

Consequences of School Dropout

The consequences of dropping out extend far beyond the educational realm. These consequences include illiteracy, unemployment, poverty, and incarceration (McCauley, 2017). Teachers and administrators face the challenge of accommodating diverse learning needs, often constrained by limited resources and support. Students with disabilities, when facing dropout, confront a future marked by limited opportunities and potential social exclusion (Papadakaki et al., 2022). Their parents grapple with concerns about their children's educational prospects and future well-being, while communities witness the ramifications of reduced educational attainment on collective growth and vibrancy. Elected officials are confronted with the social and economic implications stemming from a demographic denied the chance for educational and career advancement (McFarland et al., 2020). This widespread impact emphasizes the urgency for comprehensive strategies that involve all stakeholders in addressing this multifaceted issue.

High school dropouts not only decrease the country’s literacy rate but also contribute to higher unemployment rates (Reeves, 2021). To combat this problem, the United States government allocates funds to support the education of young citizens with special needs (U.S. Department of Education, 2022). Given the far-reaching negative impacts of high school dropout, the U.S. government remains deeply concerned about this issue. It is imperative to address the persistent dropout problem in the United States, a concern that the government continues to work on in collaboration with other global economies. Researchers have diligently

worked to provide mitigating solutions that help alleviate this widespread issue (Hussar et al., 2020; McFarland et al., 2018). In this chapter, the researcher begins by addressing the search strategy used to identify relevant studies, followed by a discussion of the theoretical framework. Key literature related to the study is then evaluated, and a summary of the chapter concludes the section.

Literature Search Strategy

The literature discussed in this chapter was sourced from online databases, notably the North Central University (NCU) Library, Google Scholar, and JSTOR, and was supplemented by resources from the Education Resources Information Center (ERIC) and Psychological Information (PsycINFO). The search process began by exploring the theoretical framework, building upon survival analysis. The focus then shifted toward identifying relevant literature sources and, ultimately, empirical studies. Keyword searches included terms, such as *dropout risk*, *high school students*, *academic motivation*, *teacher perceptions*, *predictive analysis*, and *special education*. These carefully chosen keywords helped identify literature that formed the foundation of this review.

The search strategy started broadly with general terms like “dropout risk,” “high school students,” and “academic motivation.” The search was then refined by including teacher-related terms, such as “teacher perceptions,” “teacher observations,” and “inclusive classrooms.” Reference lists from selected articles were reviewed to locate additional sources not found in the original search. Boolean operators (AND, OR) were used to combine and refine search results. A significant portion of the cited literature in this chapter was published from 1974 to 2024, with more than 90% of the sources falling between 2019 and 2024. Some earlier literature was included specifically to support the theoretical framework.

Theoretical Framework

This study's theoretical framework was rooted in a combination of survival analysis and CRISP-DM. The integration of these two frameworks allowed for examination of the temporal dimensions of high school dropout while providing a structured and methodical approach to predictive data modeling. Together, these models offer a comprehensive strategy for analyzing dropout trends and generating insights that can inform policy and practice. The aim of this framework was to support predictive modeling that identifies key variables affecting high school dropout rates and determines which of these features significantly influenced student outcomes in the state of Illinois. This combined approach helps policymakers and educators understand dropout risks and supports the design of more effective intervention strategies.

Survival Analysis Theory

The work of Graunt in 1662 led to the development of survival analysis, which is a branch of statistics that deals with analyzing the time until an event of interest occurs. It is used in various fields, such as medicine, biology, engineering, and social sciences to study the time until an event happens (C. Chen et al., 2020; Emmerson & Brown, 2021; Pan et al., 2022; Zeng, 2022). Survival analysis theory is used to estimate and interpret the survival and hazard functions from survival data. According to Emmerson and Brown (2021), the key concept in survival analysis is the hazard function, which represents the instantaneous probability of an event occurring at a given time, given that it has not occurred before that time. Survival analysis can be applied to study dropout rates in high schools in the context of the impact of COVID-19. In this scenario, the event of interest is students dropping out of high school, and survival analysis helps the researcher describe the factors influencing the timing of these events.

In recent years, machine learning algorithms have gained popularity for their ability to analyze complex datasets and extract valuable insights. Spooner et al. (2020) demonstrated that, when applied to educational data, machine learning algorithms can help identify patterns and predict outcomes, such as dropout rates. Kemper et al. (2020) also demonstrated that the use of machine learning algorithms to examine dropout rates in educational settings is a viable approach. These studies showed promising results in predicting student outcomes and identifying at-risk students who may be more likely to drop out.

Survival analysis helps assess the association of explanatory variables with survival time. Some scholars (H. L. Chen et al., 2020; Cobre et al., 2019; Martínez-Carrascal et al., 2023; Rai et al., 2021) have used survival analysis to study student dropout. The survival analysis process is used on time-to-event data. When using survival analysis, the outcome variable of interest is the time until an event occurs. The survival analysis method focuses on the duration from a starting point to an endpoint of interest. It is important to note that survival analysis presents challenges such as censored observations (Rai et al., 2021). In summary, survival analysis has proven to be an important statistical technique that can be used to identify learners at risk of dropping out (Martínez-Carrascal et al., 2023).

Understanding how the application of survival analysis helps examine the survival pattern of high school students as they encounter stress may highlight their motivational processes. For example, it may underscore the importance of interventions that enhance their expectancy for academic success, clarify the connection between academic performance and future opportunities, and amplify the perceived value of education (Wabba & House, 1974). Therefore, education stakeholders such as administrators could design improved programs to

enhance positive motivation among students with disabilities and significantly reduce dropout rates.

To mitigate high dropout rates among students with and without disabilities, properly targeted interventions are important. These interventions could focus on creating accessible and inclusive learning environments, providing tailored support systems, and fostering an understanding of how academic success leads to meaningful opportunities (Kang & Chang, 2019). By addressing these components within the framework of survival analysis, educational institutions and policymakers can work toward creating an environment that boosts motivation and persistence among students with disabilities, ultimately reducing dropout rates and increasing their chances for academic success.

Combining the insights of survival analysis and machine learning techniques benefited this study. Survival analysis provided a robust framework for modeling time-to-event data, while machine learning algorithms offered powerful tools for analyzing complex datasets and generating predictions (Kemper et al., 2020; Spooner et al., 2020). In the context of studying dropout rates in Illinois high schools during the COVID-19 pandemic, this research used survival analysis techniques to model the time until students dropped out. Researchers also incorporated variables, such as demographic information, academic performance, and socioeconomic status into machine learning models to predict which students were at higher risk of dropping out.

This study contributes to survival analysis by empirically investigating how this theory applies within the context of high school dropout. By testing the predictive relationship between demographics, geolocation, and internal school factors, the researcher aimed to validate and potentially extend survival analysis theory within this educational setting. If the hypotheses were supported, the findings would reinforce the theory's relevance in understanding motivational

factors affecting high school students, especially in relation to dropout, and would provide valuable insights for educators and policymakers. The three core concepts of the theory were discussed in this study.

The research questions and hypotheses were formulated based on the premise of survival analysis. For example, the questions examined the predictive relationship between demographic, geolocation, and internal school factors and the high school dropout rate in the state of Illinois. This line of inquiry aligns with Wabba and House's (1974) assertions on motivation. The hypotheses derived from these questions aimed to determine whether significant predictive relationships existed between these factors and high school dropout in Illinois.

CRISP-DM

CRISP-DM is a process model widely used for machine learning application development (Hayat Suhendar & Widyani, 2023). This model serves as a guiding framework that enables organizations to capitalize on their data effectively. It ensures that data are cleansed and transformed properly, leading to reliable insights. The CRISP-DM process was applied in this study to predict the high school dropout rate. CRISP-DM is a generalizable process model that provides strong guidance for advanced data science activities (Ribeiro et al., 2020). This model can be implemented with minimal training. Implementing CRISP-DM leads to better results. However, it is important to note that CRISP-DM may not be suitable for Big Data projects due to issues with velocity, veracity, volume, and variety (Saltz & Shamshurin, 2016).

CRISP-DM consists of six phases that describe the data science life cycle: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. CRISP-DM provides structured guidance for planning, organizing, and executing data science projects (Melo & de Souza, 2023; Schröer et al., 2021). In the business understanding phase,

goals and objectives are defined. In this study, the application of the CRISP-DM framework to examine the impact of COVID-19 on dropout rates in Illinois high schools began with defining the objectives and goals of the analysis. This phase required a clear understanding of the problem and its relevance to the educational landscape in Illinois (Melo & de Souza, 2023). The primary goal was to investigate how machine learning algorithms could be used to analyze the factors contributing to changes in dropout rates during the pandemic.

The data understanding phase focuses on gathering and exploring relevant datasets (Melo & de Souza, 2023; Schröer et al., 2021). These datasets should provide insights into the impact of COVID-19 on dropout rates. In this study, datasets included information on student demographics, academic performance, attendance records, socioeconomic status, and other variables that could influence dropout rates. Exploratory data analysis techniques, such as data visualization and descriptive statistics were utilized to help the researcher gain a deeper understanding of the dataset's characteristics.

The third phase is data preparation. The data preparation phase involves cleaning, transforming, and integrating datasets to ensure they are suitable for analysis (Melo & de Souza, 2023; Schröer et al., 2021). According to Kabathova and Drlik (2021), this phase is critical for ensuring the quality and reliability of data used in machine learning algorithms. In this study, the researcher addressed missing values, handled outliers, normalized variables, and performed feature engineering to enhance the predictive power of the models. A well-prepared dataset ensured a solid foundation for building effective, accurate, and reliable machine learning models.

The next stage of CRISP-DM is the modeling phase. According to Kabathova and Drlik (2021), this phase involves the application of various machine learning algorithms to build predictive models. Algorithms, such as decision trees, random forests, support vector machines,

and neural networks can be used to identify patterns and relationships within the data (Melo & de Souza, 2023; Schröer et al., 2021). The models created were designed to forecast dropout rates based on different factors related to COVID-19. This study also employed techniques like cross-validation and hyperparameter tuning to optimize model performance.

In the fifth phase, evaluation is conducted to assess model performance. The researcher evaluated machine learning models based on the goals that guided their creation (Schröer et al., 2021). In this study, the evaluation phase helped assess the performance of models in predicting dropout rates during the COVID-19 pandemic. Metrics, such as accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC) were used to evaluate model effectiveness (Kabathova & Drlik, 2021; Melo & de Souza, 2023; Schröer et al., 2021). The researcher compared different models and selected the one that best aligned with the study's objectives.

The final phase of CRISP-DM is deployment, which focuses on implementing insights derived from machine learning models into practical solutions (Kabathova & Drlik, 2021; Melo & de Souza, 2023). This phase was valuable as it helped implement insights from the model to address dropout rates in Illinois high schools. Researchers developed dashboards, reports, or decision support systems to provide stakeholders with actionable recommendations for intervention strategies (Melo & de Souza, 2023). Continuous monitoring and feedback mechanisms were essential for ensuring that deployed models remained effective over time.

Survival analysis and CRISP-DM served as the foundational framework guiding the development of the problem and purpose statements, as well as the formulation of research questions and hypotheses in the study. The selection of these theories aligned with the study's focus on understanding motivational factors influencing high school students with and without

disabilities in their decisions to continue their education or drop out (Costa et al., 2021; Wabba & House, 1974). The survival analysis theory provided a structured approach to understanding how teachers' perceptions of students' academic motivation influence their educational decisions. Specifically, the theory helped explore the links between teachers' perceptions of student motivation and the reasons students may express for lacking motivation, thereby illuminating the relationship with dropout rates among students with and without disabilities.

Review of Key Literature

A review of related literature focused on high school dropout rates and trends in the United States and examined the factors contributing to observed dropout patterns. Both individual and institutional factors were described. This section also includes the consequences of high school dropout, efforts to predict dropout, and potential remedies. A summary of the chapter concludes the section.

High School Student Dropout Rates in the U.S.

One of the prominent educational challenges in the United States involves ensuring high school graduation for every student. In a comprehensive review, Rumberger (2020) examined research findings across four dimensions: (1) the scope of the issue and its temporal trends; (2) the societal and economic consequences of high school dropout rates; (3) the underlying causes, encompassing individual and institutional factors across families, schools, and communities; and (4) potential remedies ranging from localized programmatic interventions to broader systemic reforms.

There have been fluctuations in U.S. high school dropout rates over time. Rumberger (2020) attributed these variations to a range of social, economic, educational, and policy influences. For example, dropout rates tend to rise during periods of economic hardship or social

unrest, as students face increased barriers that may hinder school attendance (Dupéré et al., 2018; Rumberger, 2020). Prior to the COVID-19 pandemic, dropout rates had declined, from 13% of all 18- to 24-year-olds in 2005 to 6% in 2018, and disparities across racial and ethnic groups had also narrowed (Bauman & Cranney, 2020).

However, during the pandemic, which disrupted educational systems globally, substantial new barriers emerged (Seymour et al., 2020). For instance, students from low-income households struggled to access the technology required for remote learning. Vulnerable populations saw a slight increase in dropout rates during this period (Seymour et al., 2020). Nonetheless, targeted interventions were implemented to address these emerging inequalities.

The reduction in high school dropout rates and improvements in graduation rates can be largely attributed to educational initiatives like early intervention programs, alternative learning pathways, and mentorship programs (Bauman & Cranney, 2020; Seymour et al., 2020). Other factors contributing to this positive trend are increased access to educational resources, improved support systems for at-risk students, and targeted interventions (Seymour et al., 2020). While recent data point to a gradual decline in high school dropout rates, the challenge remains immense.

What Factors Contribute to School Dropout?

School dropouts encompass individuals who exit the educational system before obtaining the minimum required credential. Such departures are typically instigated by external factors affecting the student, such as teen pregnancy, placement in foster care, and frequent changes in high school attendance (Farrugia & Bhandari, 2020; Kloos et al., 2023). Individuals who discontinue their high school education often cite diverse factors for their departure, encompassing school-related, family-related, and employment-related rationales. In comparison,

according to Rumberger (2020), the primary reasons indicated by tenth-graders who left school included missing too many school days, perceiving it as easier to obtain a General Education Development (GED), experiencing poor grades or failing school, disliking school, and struggling to keep up with schoolwork. However, these explanations do not uncover the fundamental root causes behind high school students dropping out of school, especially students with special needs. Thus, establishing a direct cause-and-effect relationship between any individual factor and the choice to discontinue education becomes challenging.

Two categories of factors that either contribute to or heighten the probability of high school students dropping out have been identified as follows: individual factors linked to students' traits, attitudes, conduct, and encounters, and contextual factors associated with students' families, schools, communities, and peer groups (Kloos et al., 2023). According to a review conducted by Aina et al. (2021), the likelihood of students completing or dropping out of university hinges on a combination of individual, institutional, and economic factors. The influence of these factors on the decision to drop out was moderated by a student's capacity to assimilate into the academic system.

Individual Factors

High school dropping out is linked to a range of individual factors. Rumberger (2020) notes that generally, dropout rates tend to be elevated among specific demographic groups, including males, Black and Hispanic students, immigrants, and those from language minority backgrounds. Attitudes also exert an influence on dropout rates (Benden & Lauermann, 2022). Additionally, students with lower educational and occupational aspirations are more likely to discontinue their education (Kloos et al., 2023). Several behaviors and activities serve as predictors of dropout rates, encompassing high levels of absenteeism, disruptive conduct within

the school environment, and instances of pregnancy (Mau & Li, 2018). Lastly, diminished academic performance emerges as a robust indicator of the likelihood of leaving school prematurely (Kloos et al., 2023). Taken together, these factors underscore the concept that dropout decisions are shaped by a combination of both the social and academic experiences encountered by students.

Institutional Factors

Although institutional factors undeniably play a role in students' choices to leave school, it is important to recognize that individual attitudes and behaviors are molded by the diverse environments or circumstances in which students are situated, including their family, school, community, and peer group contexts. Socioeconomic status, frequently assessed through parental education and household income, serves as a robust indicator of both academic performance and the likelihood of dropping out of school (Rumberger, 2020). Parental education plays a role in shaping students' ambitions and the level of educational support they receive, while family income affords parents the capacity to offer additional resources for their children's education (Davis-Kean, Tighe, & Waters, 2021). This includes access to higher-quality schools, after-school and summer programs, and increased educational support within the home environment.

A broad consensus by researchers stipulates that schools wield significant influence over student achievement, encompassing dropout rates (Conto et al., 2021; Mau & Li, 2018). Four categories of school attributes impact student performance: the social makeup of schools, their structural features, available resources, and their policies and practices. Beyond the roles of families and schools, communities and peer groups possess the potential to shape a student's decision to leave school (Chyn & Katz, 2021). Disparities in neighborhood attributes can offer

insights into variations in dropout rates across communities, independently of the impact of families.

In addition, factors such as interaction, course difficulty and duration, commitment, motivation, and personal family or work circumstances were found to have a less prominent role in contributing to student dropout in massive open online courses (MOOCs). This research concurred with the study by Koç et al. (2020), which revealed that the factors related to school dropout included adjustment problems at the individual, gender, family, teachers, economy, and the macrosystem level. Pov et al. (2020) also suggest that factors, such as grade retention, absenteeism, academic expectations, and engagement in private tutoring had a substantial impact on dropout rates. However, no discernible influence on dropout was observed for family- and school-related factors.

High school student dropout from school is primarily associated with family poverty, subpar academic performance, and concerns related to teachers' commitment to effective teaching within the school environment. Utilizing household survey data from thirteen low- and lower-middle-income nations from Africa and Asia with limited educational resources, Conto et al. (2021) also observed that missing or dropping out of school was linked to diminished reading and numeracy achievements. A study by Annie E. Casey Foundation & Fiester (2010), based on data from students in the United States, concluded that failure to read proficiently by the end of 3rd grade is related to higher rates of school dropout. Therefore, supporting a need to address these contributing factors to reduce the high rate of high school dropout in the U.S.

Ginevra et al. (2021) investigated the correlation between various personal attributes of teachers (such as gender, age, years of professional experience, and the grade level they teach), characteristics of students with disabilities (including disability type and gender), and contextual

factors (specifically, the type of information available in the profiles of students with disabilities). They revealed that secondary school teachers tended to hold more unfavorable attitudes regarding the social acceptability of students with disabilities. Moreover, they found that teachers exhibited more negative attitudes regarding academic performance when dealing with male students and students with intellectual disabilities, and more negative attitudes concerning social acceptability when working with students exhibiting behavior problems (Ginevra et al., 2021). Notably, the study also indicated that teachers' attitudes toward students were positively influenced when the information provided in students' profiles emphasized their strengths. These findings highlight the nuanced nature of teachers' attitudes and underscore the significance of the information provided about students with disabilities. Emphasizing strengths in students' profiles emerged as a potential strategy to positively influence teachers' perceptions, indicating the importance of how student information is presented and the need to accentuate positive attributes to foster more supportive and inclusive learning environments.

A study that delved into the perceptions of middle school mathematics teachers regarding their teaching approaches with students with disabilities (SWDs) revealed that teachers' perceptions played a pivotal role toward the student's achievements. Myers et al. (2021) examined these practices in both traditional face-to-face instruction prior to the COVID-19 pandemic and during the subsequent eLearning period. Through a survey, teachers' utilization of research-backed techniques and the factors influencing their instructional methods were explored (Myers et al., 2021). The study's outcomes revealed that teachers reported relatively consistent classroom practices across both modes of instruction. Additionally, teachers expressed challenges in providing accommodations and implementing small group instruction during eLearning. They highlighted barriers to eLearning associated with student engagement and

instructional planning and delivery. Finally, the study discussed its findings, constraints, and implications for future research, as well as considerations for the preparation of both pre-service and in-service educators. On the contrary, Lorenzo-Lledó et al. (2020) noted that teachers believed that students with disabilities should attain the same skills as their peers and that they did not find it challenging to instruct them.

A study administered to high school general education teachers to assess their perceived expertise and competence in instructing students with significant disabilities revealed a need for training and experience in special education. The study by Kroesch and Peeples (2021) encompassed a variety of subject areas, years of teaching experience, and levels of educational training. The findings from this study corroborated with previous and recent research findings that underscore the difficulties associated with ensuring a top-tier inclusive learning environment for students with significant disabilities (Hagaman & Casey, 2018). The implications drawn from this study emphasized the necessity for general education teachers to receive training and gain experience in special education topics to enhance their ability to effectively educate students with significant disabilities within inclusive educational settings.

Motivating students to stay in high school and complete their education is a multifaceted challenge. Understanding the motivations of these students in the context of high school dropout is critical for developing effective interventions and support systems. Students with disabilities have unique motivations that drive their decisions regarding school attendance and completion (Hardré et al., 2008). These motivations can include academic success, the prospect of a fulfilling career, social connections, personal growth, or the desire to prove themselves (Ginevra et al., 2021). Academic challenges play a significant role, with increased difficulties leading to decreased motivation when students feel overwhelmed by coursework, face academic failures, or

perceive a lack of relevance in their education (Hardré et al., 2008; Papadakaki et al., 2022). Emotional and social factors also come into play, with factors, such as stigma, bullying, and social isolation eroding motivation, while positive social interactions, supportive friendships, and a sense of belonging can boost it.

Family dynamics, self-perception, future aspirations, and perceived relevance of education also shape motivation. Supportive teachers, mentors, and personalized learning can enhance motivation (Kroesch & Peeples, 2021; Myers et al., 2021), while early intervention, positive reinforcement, and peer support can play crucial roles in maintaining and strengthening students' motivation to complete their education (Chitsamatanga & Rembe, 2020; Gausel & Bourguignon, 2020; Li & Carroll, 2020; McFarland et al., 2020). Ultimately, motivation among high school students with disabilities is influenced by a complex interplay of academic, social, emotional, and individual factors, and addressing these comprehensively is essential to support their educational success (Foreman-Murray et al., 2022). It is therefore clear that both individual and institutional factors play a critical role in determining high school students' educational completion or dropping out.

Consequences of High School Dropout

Dropping out of school carries significant economic and societal repercussions, affecting both the individuals who leave education prematurely and the entire nation. Initially, dropouts encounter challenges in securing employment, as government statistics from October 2015 indicate that over 37% of individuals aged 16 to 24 who had dropped out of school before 2015 were unemployed compared to those who completed high school (Rumberger, 2020). Furthermore, even if they managed to secure employment, dropouts earned considerably lower wages than their counterparts who had completed high school. In 2005, the median annual

income of high school dropouts was nearly 25% less than that of high school graduates (Snyder et al., 2018). Additionally, research by Koç et al. (2020) exploring students' experiences after school dropout established that individuals who left school prematurely encountered difficulties entering the job market or found themselves in low-status occupations. These individuals expressed regret over their decision to drop out and believed that their lives would have been more favorable had they remained in school. A study by Li and Carroll (2020) revealed that students belonging to equity groups often exhibited lower academic performance and were more inclined to contemplate dropout, primarily driven by health and financial concerns. This underscores the significance of implementing comprehensive initiatives to provide support for these students.

Studies have shown that high school dropouts, including those with and without disabilities, are at a higher risk of becoming involved in criminal activities and subsequently being incarcerated. Incarceration can further limit their future opportunities and perpetuate a cycle of disadvantage (McCauley, 2017). High school dropout is associated with unemployment, low income, poor healthcare access, illiteracy, poor living conditions, and indulgence in criminal activities (McFarland et al., 2020; Montes & Mendes, 2021). These human wellness constraints have an overall impact on the mental health of the affected individuals.

High school dropouts often have lower levels of civic engagement and community involvement. This disengagement can have negative consequences for both the individual and society in terms of participation in community life and decision-making processes (Gausel & Bourguignon, 2020). Overall, dropping out of high school can lead to a reduced quality of life for individuals, with limited access to opportunities for personal and professional growth (Lee & Polachek, 2018). It can also contribute to feelings of social isolation and marginalization

(Montes & Mendes, 2021). Due to limited economic prospects, many high school dropouts, especially those with disabilities, may rely on government assistance programs for their livelihood (Lee & Polachek, 2018). This places a burden on government resources and social services.

Predicting Students' Dropout Risk

According to the National Center for Education Statistics (2023), in 2021, there were 2.0 million status high school dropouts between the ages of 16 and 24. The high school dropout rate is a persistent challenge in online learning, but the random forest regression model stands out as it surpasses other algorithms in its ability to predict student dropout and pinpoint actionable factors for supporting students in their learning journey (Jimenez et al., 2023). This model offers unique insights tailored to diverse student segments, allowing for the customization of student retention initiatives. In their study, Wen et al. (2020) introduced a novel Convolutional Neural Network (CNN) model designed to take into account the localized correlations in learning behaviors for predicting dropout in massive open online courses (MOOCs). This model could provide early and temporal dropout predictions, contingent upon the accumulation of sufficient data.

In addition to ensemble and deep learning methods, multiple linear regression (MLR) has long served as a foundational tool in educational research for modeling dropout behavior. Its interpretability and simplicity make it ideal for identifying and quantifying the direct effect of individual predictors on student outcomes. Studies such as those by Ahmed et al. (2025) have demonstrated the value of MLR in estimating the influence of socioeconomic status, attendance, and academic performance on dropout risk. While MLR may not capture complex interactions or nonlinear patterns as effectively as advanced machine learning models, its transparency offers

critical insights into the magnitude and direction of specific factors, information that is often directly applicable to policy development.

Meanwhile, the XGBoost model has gained traction in educational data mining due to its high accuracy, efficiency, and capability to handle large-scale, noisy datasets. XGBoost builds on the advantages of decision trees and gradient boosting by incorporating regularization, handling missing data, and optimizing computation. In comparative studies, XGBoost often outperforms traditional algorithms in predicting academic outcomes, including dropout (Deleña et al., 2025). Its ability to rank feature importance and model intricate nonlinear relationships makes it especially valuable in uncovering hidden patterns in student data, which can inform targeted interventions across diverse educational settings.

Drawing upon a temporal prediction framework, Xing and Du's (2019) research advocated for the utilization of a deep learning algorithm for crafting a dropout prediction model capable of estimating individual student dropout probabilities. Leveraging the capabilities of deep learning, this approach not only resulted in more precise dropout prediction models in contrast to baseline algorithms but also introduced a method for customizing and prioritizing interventions for students at risk by leveraging individual dropout probabilities.

Several factors have been identified as predictors of student dropout. These factors can be categorized into individual, academic, socio-economic, and environmental variables. Individual factors include students' demographic characteristics, prior academic performance, and behavioral issues (Mughal, 2020; Rumberger, 2020; Xing & Du, 2019). Academic factors encompass course performance, attendance, and test scores. Socio-economic factors involve family income, parental education, and access to resources. Environmental factors include the schooling environment, teacher-student relationships, and peer interactions (Aldowah et al.,

2020; Conto et al., 2021; Wen et al., 2020). In the quest to predict student dropout, statistical models have been widely employed. These models use historical data to identify patterns and relationships among various predictors. Among the statistical methods commonly used are logistic regression, multiple linear regression, decision trees, and machine learning algorithms, such as neural networks, random forests, and XGBoost.

Dropout Rate and Machine Learning Techniques

Numerous students did not keep up with their studies due to the consequences of the COVID-19 pandemic (Colpo et al., 2024; Lichand et al., 2022; Shuja et al., 2022). Some common factors are causing a large number of high school and college students to drop out of school (Cassel, 2003). The low literacy and involvement in alcohol and drugs cause these dropouts to get incarcerated (Gausel & Bourguignon, 2020; Hussar et al., 2020; McFarland et al., 2018).

Hossain et al. (2022) conducted an effective study on the dropout rate of university students in Bangladesh. These researchers investigated the dropout rate of university students using machine learning models that include random forest, support vector machine, and logistic regression. These researchers compared these machine learning models to find out which one would best predict the dropout rate of university students.

Using multiple linear regression to analyze data makes it easy to understand and do the interpretation. This model is flexible and adaptable. However, this linear regression model is sensitive to outliers and noise and prone to overfitting and underfitting. While multiple regression models help to analyze the influences of predictor variables on the response variable, the analysis of some huge datasets not done properly can lead to false conclusions (Liu et al., 2022).

Using random forest helps to build a forest with an ensemble of decision trees. This model is an easy-to-use machine learning algorithm that reduces overfitting in decision trees and helps to improve accuracy. The flexibility of random forest to both classification and regression problems makes it suitable to address a broader range of problems. Random forest model works well with both categorical and continuous values and automates missing values present in the data. However, it is important to point out that random forest does not provide complete visibility into the coefficients as linear regression. Random forest model is computationally intensive for large datasets (Langsetmo et al., 2023).

The XGBoost model is a powerful and efficient machine learning algorithm that extends gradient boosting techniques with regularization, parallel processing, and optimized tree learning (Inoue et al., 2020). XGBoost excels at modeling nonlinear relationships and complex feature interactions, making it particularly well-suited for high-dimensional and structured datasets commonly found in educational data analysis. Its strength lies in its ability to iteratively correct prediction errors made by previous models through a boosting process, thereby improving accuracy with each additional tree (Inoue et al., 2020). This makes XGBoost especially effective in dropout prediction tasks where patterns may not be easily captured by linear assumptions. XGBoost also incorporates robust mechanisms to handle missing data, multicollinearity, and class imbalance, offering an edge over traditional models when dealing with real-world educational datasets that are often incomplete or noisy (Jinbo et al., 2025). It provides tools for feature importance ranking, enabling researchers to identify key factors influencing dropout risks across varied school environments. Despite its high performance, XGBoost is more computationally demanding than simpler models and can be sensitive to hyperparameter tuning (Tarwidi et al., 2022). Additionally, its interpretability is limited compared to models like

multiple linear regression, requiring the use of supplementary tools such as SHAP (SHapley Additive exPlanations) to fully understand model outputs (Chen & Guestrin, 2016).

Traditional methods used by school districts to examine and mitigate school dropout were not as effective as modern methods. Traditional methods examine the dropout issue through a lens focusing primarily on student variables, which fail to account for the interaction between students and their environments (Allensworth & Easton, 2007; Balfanz et al., 2007; Neild, 2009). For instance, Early Warning Systems have shown some reliable results concerning the identification of students at risk of dropping out. However, they make use of data that may not capture the full range of factors that contribute to dropout risk.

Machine learning algorithms use data to identify at-risk students early so that dropout can be prevented. These algorithms have the potential to predict student dropout accurately by harnessing extensive data and advanced analytical techniques. Machine learning models have been shown to predict student dropouts accurately (Song et al., 2023). These models can identify students who are more susceptible to dropping out, as they examine different factors and intricate patterns. Machine learning algorithms allow significant advances in the prevention of school dropout and help address its economic and social impacts (Segura et al., 2022).

In this study, multiple linear regression, a machine learning algorithm used for supervised learning, was used to predict high school dropout rates based on the demographic, geolocation, and other internal factors of the schools. This statistical method is used to analyze the relationship between a single response variable and multiple predictor variables. This model is used to determine the influence of one or more independent variables on the response variable. Multiple linear regression aims to predict the value of the single dependent variable using predictor variables whose values are known (Adasme et al., 2023). This model helps identify

which predictor variables have a strong correlation to the response variable. In addition, it helps identify outliers and anomalies.

Despite the usefulness of this model, there are some drawbacks associated with its implementation. These drawbacks include susceptibility to outliers, limited flexibility, overfitting, assumptions of multicollinearity, and assumptions of linearity. This model assumes a linear relationship between the response variable and predictor variables. When the relationship between the variables is not linear, this assumption is not met. It is more complicated to deal with nonlinear relationships, as more complex models may be required to capture their nuances.

According to Q. Chen and Lee (2021), the random forest regression model, a machine learning algorithm used for supervised learning, was also used to predict high school dropout rates based on the demographic, geolocation, and other internal factors of the schools. This model is an ensemble learning technique that is especially useful for addressing classification and regression problems.

Random forest helps address the problem of overfitting and high variance in decision trees (Jimenez et al., 2023). This model is a data-robust algorithm that handles different types of data, combines the results of different decision trees, and performs better than many learning algorithms (Dass et al., 2021). Random forest is robust to outliers and multicollinearity.

Despite the usefulness of this model, there are some drawbacks associated with its implementation. These drawbacks include difficulty in describing relationships within data, computational complexity, memory usage, and overfitting. Random forest may be too slow and ineffective in making predictions due to an increased number of trees. In addition, the random forest model uses substantial memory when working with large datasets.

This study aimed to predict the school dropout rate based on the demographic, geolocation, and other internal factors of the schools. Choosing multiple linear regression and random forest to conduct this study was appropriate, as both models can be used to predict the response variable based on predictor variables. The response variable in this study was continuous. In addition, multiple linear regression is flexible and adaptable and makes interpretation straightforward. Random forest works well with both categorical and continuous values and handles missing values in the dataset. The dataset used in this study contained both categorical and continuous variables.

According to Jinbo et al. (2025), the XGBoost regression model, which is also a machine learning algorithm used for supervised learning, was included in this study to predict high school dropout rates based on demographic, geolocation, and internal school-level factors. XGBoost is an advanced ensemble method that builds upon gradient boosting principles, where decision trees are constructed sequentially, and each subsequent tree corrects the errors of its predecessor. This approach allows XGBoost to produce highly accurate and robust models, especially when dealing with large and complex datasets. XGBoost has become a popular choice in predictive analytics due to its ability to handle missing values, multicollinearity, and nonlinear relationships more effectively than traditional models. The XGBoost model also provides feature importance scores, which can help identify the most influential predictors of high school dropout rates.

Despite the advantages of the XGBoost model, there were limitations associated with its implementation. These included the need for careful hyperparameter tuning, which was time-consuming and required technical expertise to avoid overfitting or underfitting. Additionally, XGBoost was less interpretable than simpler models like multiple linear regression, as it did not provide clear coefficient estimates for each predictor. Instead, it required additional tools such as

SHAP values to explain individual predictions. Furthermore, XGBoost was computationally demanding, especially with very large datasets or when many boosting rounds were required. However, its ability to model complex interactions and nonlinearities made it a valuable addition to this study's modeling framework.

Teachers play a crucial role in identifying students at risk of dropping out. Their observations and perceptions of students' academic motivation can provide valuable insights into potential dropout risks. Several studies have explored the impact of teachers' perceptions on dropout prediction. For example, the study by Kroesch and Peeples (2021) examined teachers' perceived knowledge and capabilities related to teaching students with disabilities, highlighting the importance of teacher awareness in dropout prevention.

When considering high school students with disabilities, additional factors come into play. These students often face unique challenges, including learning disabilities, physical disabilities, and behavioral issues (Myers et al., 2021). Research specific to this population has shown that factors, such as the type of disability, access to support services, and teacher–student relationships can significantly influence dropout rates (Li & Carroll, 2020). One emerging trend in dropout prediction and prevention is the personalization of interventions. Recent studies have emphasized the need for tailored strategies based on individual student profiles (Coussement et al., 2020). This approach involves using predictive models to identify students at risk and then implementing targeted interventions based on their specific needs and motivations.

In contrast, a study by Košir et al. (2023) on the application of school attachment factors as a strategy against school dropout revealed that several socioeconomic constraints remained a confounding factor in high school dropout. The study findings indicated that although interest in learning and positive inclinations about the school had a critical impact on students' interaction

with their school, positive attitudes toward the teaching fraternity were the greatest determinant of high school students' attachment to school. The findings suggested a need to upgrade and strengthen the role and efficacy of teacher–student relationships to arrive at better educational outcomes for high school students.

Remedies to High School Dropout

Rumberger (2020) examined research findings relating to remedies for high rates of high school dropout and revealed potential ways of addressing this challenge. The remedies included a range of proposals spanning from localized programmatic interventions to more extensive systemic solutions within educational institutions. Localized programmatic interventions helped develop targeted and effective strategies to alleviate high school dropout rates (Ramsdal & Wynn, 2022). Specific communities or schools benefited from such initiatives since unique challenges were addressed in a tailored manner. Ramsdal and Wynn (2022) argued that such programs could include efforts to ensure readmission of students who had dropped out of school. For instance, mentorship programs that paired students at risk of dropping out were proposed by W. Y. Chan et al. (2020). They argued that pairing at-risk students with mentors could increase motivation to persist in school. Ramsdal and Wynn (2022) believed, based on their observations, that such interventions provided special guidance, encouragement, and support needed to keep learners engaged in schools. Also, providing after-school programs that improved engagement and skill development, such as extracurricular activities, could reduce dropout rates and increase completion rates (Rumberger, 2020).

Counseling and therapy services also helped alleviate high school dropout rates. Having access to counseling services within schools helped address underlying problems that influenced students' dropout decisions (W. Y. Chan et al., 2020; Ramsdal & Wynn, 2022; Rumberger,

2020). For example, mental health challenges and family and social problems, if left unaddressed, could increase dropout risk. Providing such services in high schools increased access and quality, serving as a potential remedy to the challenge.

In addition to localized solutions, systematic institutional interventions also helped address dropout. These interventions encompassed institutional-level changes aimed at creating more supportive and inclusive learning environments (Rumberger, 2020). Early intervention strategies that targeted vulnerable learners before they reached the crisis stage were found useful (Ramsdal & Wynn, 2022; Rumberger, 2020). Collaborations with stakeholders to build resources supporting students were also cited as effective ways to reduce dropout (W. Y. Chan et al., 2020). This assertion aligned with the need to offer support systems, including mental health services, counseling, and therapy, through partnership initiatives from high schools themselves.

Data-driven decisions significantly helped address high school dropout. Kearney and Childs (2023) demonstrated that effective use of data analytics could track students' attendance, performance, and behavioral indicators, and ultimately predict outcomes such as dropout likelihood. Xing and Du's (2019) study argued that it was effective to utilize a deep learning algorithm for crafting a dropout prediction model capable of estimating individual student dropout probabilities. These algorithms offered a more accurate approach to remedy the problem of high school dropouts.

Summary of Empirical Review

This section presented a review of empirical literature aligned with this study's objectives and methodological approach. This study aimed to answer the following research questions:

1. To what extent, if any, did demographic, geolocation, and other internal school factors predict the high school dropout rate in the state of Illinois?

2. Which predictor variables significantly influenced the prediction of the high school dropout rate in the state of Illinois?

The review revealed that school dropout rates were significantly associated with several factors, including school location, learner characteristics, and environmental variables. These factors were found to be predictable with a significant level of accuracy using robust machine learning techniques, in contrast to traditional statistical approaches.

Summary

Multiple researchers have examined factors that caused students to leave school without completing their high school education (Irwin et al., 2021; McFarland et al., 2018). The findings from these studies revealed that high school dropouts experienced negative consequences in different areas. However, little was known about conditions that triggered high school dropout (Dupéré et al., 2018). Wan (2022) pointed out that there remained a gap in the literature relating to the motivation for the decision to drop out of school. If dropping out of school was not addressed, economic and social consequences would continue to intensify steadily (Lee & Polachek, 2018). Therefore, this issue needed to be examined carefully to provide school officials with insights that would help them reduce the dropout rate significantly.

The issue of high school dropout remained a persistent and pressing concern in the United States and globally. The magnitude of the problem, with over 1.2 million students leaving high school each year in the United States alone, highlighted the urgent need for effective interventions (Miller, 2011). High school dropout not only hampers the prospects of individuals who leave school prematurely but also has far-reaching social and economic consequences (Lee & Polachek, 2018). Education is widely recognized as a catalyst for societal development,

impacting areas, such as crime rates, overall health, and civic engagement (Chitsamatanga & Rembe, 2020; Gausel & Bourguignon, 2020).

Of particular concern is the disproportionately high dropout rate among students with disabilities, who are more likely to leave high school before graduating (McCauley, 2017). The consequences for these students can be severe, including increased incarceration rates and reduced earning potential (McCauley, 2017; U.S. Bureau of Labor Statistics, 2021). Despite extensive research in this area, including studies examining various factors contributing to dropout, such as social, behavioral, and academic challenges, the issue remains complex and multifaceted.

The motivation behind students' decisions to drop out is a critical aspect of this problem, and it is here that this study introduced a novel approach. By employing survival analysis, which helps understand the relationship between the dependent variable and the predictor variables, and CRISP-DM, which helps capitalize on data with effectiveness, this study sought to examine the prediction of high school dropout rates based on the demographic, geolocation, and other internal factors of the schools.

This quantitative exploratory study aimed to develop and evaluate a predictive model for dropout risk among high school students based on these variables (Koç et al., 2020; Szlyk, 2020). By doing so, the study aspired to provide valuable insights that could inform targeted interventions and support systems to reduce dropout rates and improve the prospects of these students.

In essence, the complex issue of high school dropout, especially among students with disabilities, demands multifaceted solutions. Addressing this problem goes beyond the realm of education, affecting various facets of society. Some remedies were proposed by previous

researchers, including interventions spanning from localized programmatic solutions to more extensive systemic changes within educational institutions (W. Y. Chan et al., 2020; Ramsdal & Wynn, 2022; Rumberger, 2020). However, a more recent focus on data-driven decisions to address the problem of high school dropouts has not been explored extensively, despite its immense potential (Kearney & Childs, 2023; Xing & Du, 2019). These studies demonstrated how effective data analytics can help track students' school attendance, performance, behavioral indicators, and ultimately predict outcomes such as dropout risk.

As the United States government continues to invest in education and recognizes the far-reaching consequences of dropout, research efforts like this one contribute to understanding the problem and offer pathways toward effective solutions. This study's theoretical framework, guided by survival analysis and CRISP-DM, provided a promising approach to uncovering the motivations behind dropout decisions, ultimately paving the way for tailored interventions and improved outcomes for high school students.

Chapter 3: Research Method

Predictive modeling helps organizations analyze data patterns to predict future events (Subho & Chowdhury, 2022). Despite numerous studies examining dropout rates among high school students, relatively few have leveraged the power of machine learning algorithms to predict dropout risk for this demographic (Skittou et al., 2024; Soland et al., 2020; Uldall & Rojas, 2022). Traditional methods of dropout prediction have largely relied on statistical analyses and socio-demographic profiling; however, these approaches often fail to capture the complex and multifaceted nature of dropout behavior.

Recent advancements in machine learning have opened new avenues for more accurate and dynamic prediction models. For instance, a study conducted in California utilized random forest algorithms to identify key predictors of dropout among high school students, significantly improving prediction accuracy over traditional methods (Smith & Jones, 2020). Similarly, research in Texas applied support vector machines to educational data, demonstrating the potential of machine learning to uncover hidden patterns and risk factors not easily identified through conventional analyses (Brown et al., 2019). In addition, machine learning techniques have been successfully applied at various educational levels and contexts. In higher education, researchers used neural networks to predict college dropout rates, revealing critical insights into student engagement and performance metrics (Lee & Kim, 2018). Another notable study in the context of vocational training programs in New York employed gradient boosting machines, resulting in a robust model for early identification of at-risk students (Doe et al., 2021).

Booker et al. (2010) conducted a pivotal study to estimate the effects of attending a charter high school on the likelihood that a student would drop out, complete high school, and attend college. The findings of these researchers highlighted a significant gap in the literature,

noting that less attention had been given to examining dropout rates specifically in the context of charter high schools. The conclusion of that pivotal study was made despite the growing body of research focusing on the causes of dropout risk in general. For instance, factors such as socioeconomic status, family background, and academic performance are well documented as influencing dropout rates (Bowers & Sprott, 2012; Rumberger & Lim, 2008).

Comprehensive studies that analyze these factors using advanced methodologies like machine learning remain sparse. Paksi et al. (2023) emphasized that while there is a wealth of research on the causes of dropout risk, few studies approach the problem in a multifaceted manner that captures the complexity of the issue. Machine learning algorithms offer a promising solution to this problem by enabling the analysis of large datasets and the identification of intricate patterns and relationships that traditional statistical methods might overlook (Xu & Jaggars, 2014).

In a study by Kotsiantis et al. (2007), various machine learning techniques, including decision trees and neural networks, were employed to predict student dropout in Greek higher education. Their findings demonstrated the potential of these techniques in providing accurate predictions and insights. Similarly, Herzog (2006) applied data mining methods to predict student attrition in higher education, further validating the efficacy of machine learning approaches in educational contexts.

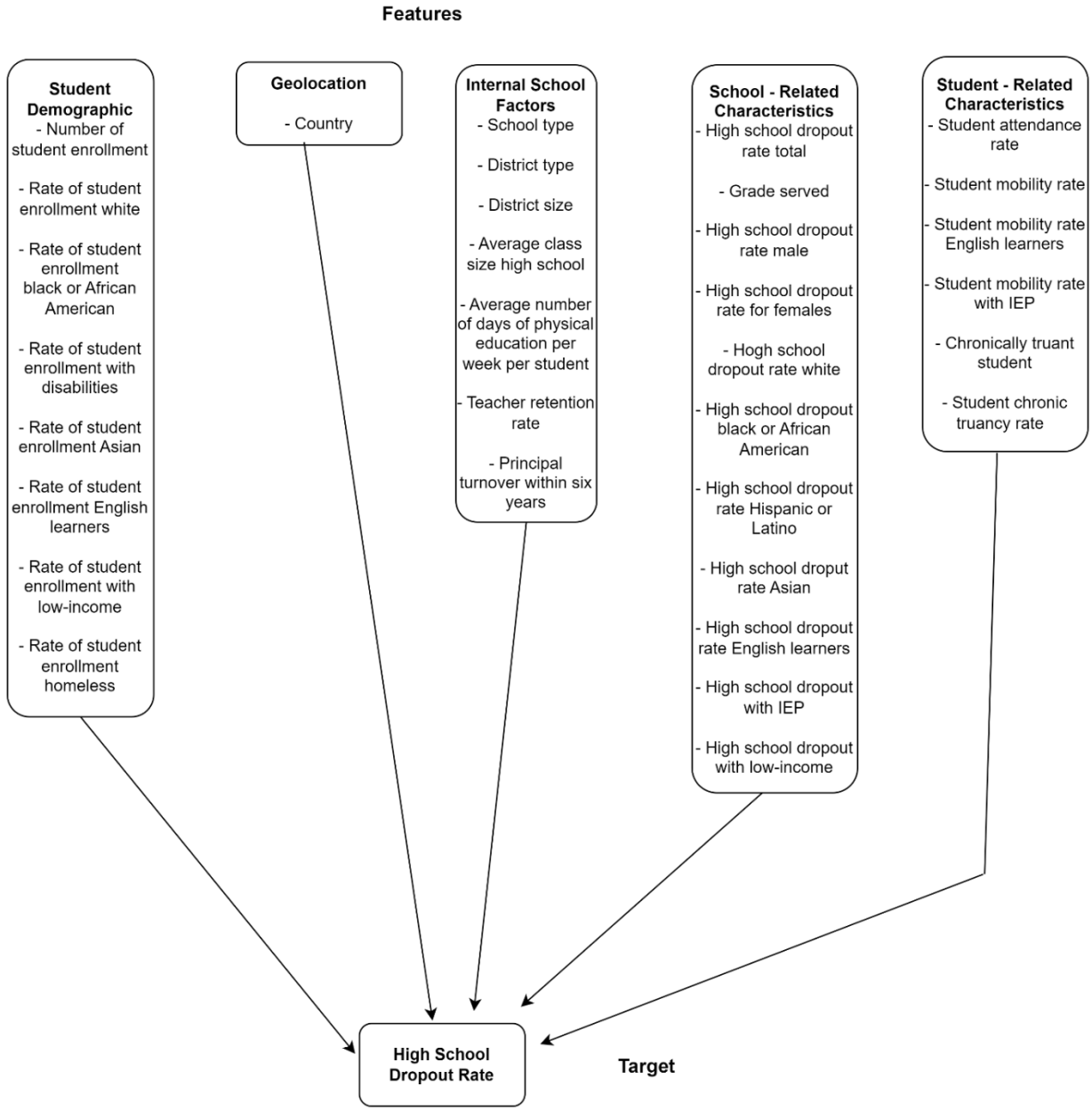
In this chapter, the use of machine learning algorithms is discussed to examine the response variable, which is high school dropout rates. By integrating multiple predictors and utilizing sophisticated analytical tools, this research offers a more nuanced understanding of the factors contributing to dropout rates, particularly in the context of charter high schools. This

approach builds on the foundational work of Booker et al. (2010) and expands the scope of dropout research by incorporating advanced machine learning methodologies.

The purpose of this quantitative study was to use specific predictive modeling techniques, discussed later, to predict high school dropout rates based on student demographic, geolocation, and other internal factors and characteristics of high schools. Numerous studies revealed that internal school factors, including class size, class schedule, and school environment, contribute to school dropout (Boyaci, 2019; Narvaez & Gomez, 2023). Berg and Nelson (2016) conducted a study and showed that school characteristics, including school leadership, building, and learning environment, had an effect on high school dropout. Below is a conceptual model that graphically represents features and the target (see Figure 1).

Figure 1

A Conceptual Model that Graphically Represents Features and the Target

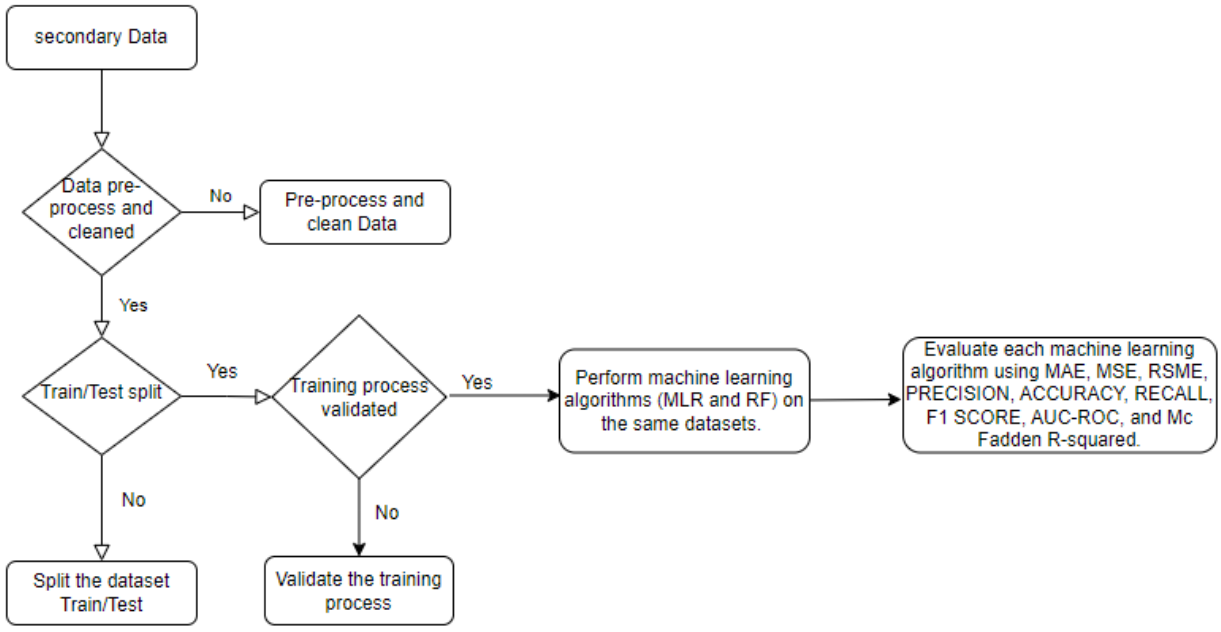


In this Chapter, a description of the research methodology and design, as well as the alignment with the problem statement, the purpose of the study, and the research question, is provided. The population, sample, and data collection process are described as well. The last portion of this chapter presents operational definitions of variables, study procedures, data

analysis, assumptions, limitations, delimitations, ethical assurances, and a summary. The methodology used to conduct this study is summarized in Figure 2.

Figure 2

Overview of the Methodology Plan



Research Methodology and Design

This study applied a quantitative, non-experimental design to build a predictive model for the high school dropout rate based on demographic, geolocation, and other internal factors of the schools. A multiple linear regression model, a supervised machine learning technique, was used to predict the dropout rate based on those variables (Kim et al., 2021). Considering the recommendations of Evans et al. (2010), predictive modeling, particularly multiple linear regression, random forest, and XGBoost, was deemed appropriate for this study, as the research question focused on examining the predictive relationship among multiple variables.

The multiple linear regression model assumed a linear relationship between the independent variables and the dropout rate; therefore, the data were evaluated for linearity through exploratory data analysis, including residual plots and correlation matrices. The random forest model, a regression-based non-parametric approach, did not assume linearity and served as a complementary method capable of capturing complex, potentially nonlinear relationships among variables. The XGBoost regression model, like the random forest, is a non-parametric machine learning technique that does not assume linearity between the independent variables and the dropout rate. Instead, it uses a gradient boosting framework to iteratively improve predictions by minimizing error through sequential decision trees. This model is particularly effective at capturing nonlinear relationships, complex interactions, and variable dependencies that may be overlooked by simpler models.

As part of this study, XGBoost was implemented alongside multiple linear regression and random forest to provide a high-performance, fine-tuned prediction model. Model tuning involved adjusting hyperparameters, such as learning rate, maximum tree depth, and number of boosting rounds to optimize predictive accuracy and prevent overfitting. The inclusion of XGBoost enhanced the robustness of the analytical approach and supported comprehensive model comparison across linear and nonlinear predictive techniques. The use of these models enabled comparison of predictive performance and interpretation, balancing model accuracy with explanatory insight. Few studies have examined the learning achievement of students based on student mobility rate, student attendance rate, and rate of students with Individualized Education Program (IEP; Brendan, 2020; Grigg, 2012).

Quantitative research designs are divided into two major types: experimental and descriptive non-experimental (Novosel, 2022). Given that no intervention was introduced to

influence the outcome of this study, the experimental design was excluded. A non-experimental design was appropriate for this study, as it aimed to examine the predictive relationship among variables at a particular point in time (Al Kuhayli et al., 2019; Novosel, 2022). Because the study employed a non-experimental (observational) design using administrative data from Illinois high schools (2017–2022), the analysis necessarily focused on associations and prediction rather than causal effects. Without random assignment or controlled interventions, unobserved confounding could not be ruled out, and the models did not support counterfactual claims. Consequently, all findings were framed as predictive relationships between school-level characteristics and dropout rates. Data manipulation was limited to reproducible preprocessing, cleaning, encoding, imputation, transformation, and principled variable reduction, rather than any treatment assignment or experimental manipulation. This lack of experimental control also implied that generalization depended on the stability of the observed data-generating process. The study therefore emphasized cross-validated, held-out evaluation to guard against overfitting.

This design choice affected data handling in several concrete ways. First, the study prioritized transparent, rule-based retention of variables over outcome-driven choices. Features with more than 50% missingness were removed to avoid unstable imputations. Potential redundancy was screened via the correlation matrix (flagging pairs with $|r| \geq .70$) and by Variance Inflation Factor (VIF), with 7.5 used as a pruning threshold (10 recognized as severe). On this basis, *year* was dropped because it overlapped with *covid_period* and exhibited a non-trivial VIF (≈7.73), and *chronically_truant* was removed because it tracked the same construct as *chronic_truancy_rate*. Missing values were then imputed using simple, robust procedures that preserved the original scale of measurement, medians for numeric fields and the mode for the encoded categorical field, thereby avoiding model-dependent imputations that could complicate

interpretation. Outliers were inspected but not winsorized or removed in bulk; robustness was addressed through model choice (tree-based ensembles) and post-fit residual diagnostics. To prevent data leakage, all transformations were fit on training folds only and then applied to validation and test data, ensuring that evaluation metrics reflected genuine generalization.

The observational setting also raised the risk of hyperdimensionality, too many, often correlated predictors relative to the effective information in the data, so the study used a layered reduction strategy that preserved interpretability. At the filter stage, variables were eliminated based on the missingness and collinearity rules noted above. At the embedded stage, the linear family employed regularization (Ridge, Lasso, and Elastic Net) during cross-validated tuning to shrink unstable coefficients and, where appropriate, set them to zero, improving parsimony without inventing latent components. For the non-parametric models, complexity was controlled through model-based constraints: Random forest used limits on tree depth, minimum samples per split/leaf, and the number of trees. XGBoost used learning rate, maximum depth, and subsample/colsample parameters to regularize the boosted ensemble. Across all models, 10-fold cross-validation and a held-out test set provided out-of-sample checks on R^2, *MAE*, *MSE*, and *RMSE*, so that dimensionality choices were validated empirically rather than tuned to the training data.

Principal components analysis (PCA) was considered as an additional dimensionality-reduction tool but was not adopted in the main analyses because it transforms predictors into orthogonal components that are difficult to map back to concrete school levers, reducing policy interpretability. Where PCA was used for sensitivity analysis, components were derived from centered and scaled predictors on training folds only, with the number of components chosen to preserve a high proportion of variance (e.g., 90%–95%), and downstream models were re-

estimated on those components. The primary results nevertheless relied on regularization and filter-based pruning, which maintained a direct link between predictors and actionable factors.

In sum, the lack of experimental control meant that inference was predictive and associational, not causal. This limitation motivated a conservative data manipulation strategy: explicit removal thresholds (i.e., >50% missingness, $|r| \geq .70$, VIF ≥ 7.5), median/mode imputation fit within cross-validation, regularization for feature selection and stability, and model-based complexity control in tree ensembles. These choices reduced hyperdimensionality while preserving interpretability and the integrity of out-of-sample evaluation.

The observational setting also heightened the risk of hyperdimensionality, too many, often correlated predictors relative to the effective information in the data, leading to variance in inflation and overfitting. The study responded with a layered strategy that combined design-stage pruning, statistical screening, and model-based regularization while preserving interpretability. At the design stage, features with extensive missingness and near-duplicates were removed, and derived indicators were used selectively (e.g., retaining *covid_period* while dropping *year*) to reduce redundancy. Statistically, the study used pairwise correlations and VIF to identify and remove collinear predictors before modeling. To make variable retention transparent and reproducible, predictors were screened against prespecified criteria rather than outcome-driven choices.

First, any feature with more than 50% missingness across the pooled 2017–2022 panel was removed to avoid unstable imputations; this included several subgroup dropout rates and sparsely reported subgroup enrollment counts (e.g., gender- and race-specific rates). Second, potential multicollinearity was evaluated in two steps: pairs with $|r| \geq .70$ in the correlation matrix were flagged for review, and variables with VIF ≥ 7.5 were treated as candidates for

pruning (with VIF ≥ 10 recognized as a conventional “severe” threshold). On this basis, *year* was dropped because it was largely redundant with *covid_period* and showed a non-trivial VIF (~7.73), while *chronically_truant* was removed due to substantive and statistical overlap with *chronic_truancy_rate*. After these exclusions, the retained predictors exhibited acceptable correlations and VIFs under the stated thresholds. Missing values were imputed using medians for numeric variables and the mode for the encoded categorical field, preserving scale and interpretability for downstream modeling.

Within the modeling step, the linear family employed regularization (Ridge, Lasso, and Elastic Net) to shrink or zero out unstable coefficients, improving parsimony without inventing latent components. For the non-parametric models, random forest complexity was controlled through limits on depth, leaf size, and the number of trees, and XGBoost used learning-rate control, shallow depth, and subsampling to regularize the boosted ensemble. Across all models, 10-fold cross-validation and a held-out test set provided out-of-sample checks on error (*MAE*, *MSE*, *RMSE*) and fit (R^2, adjusted R^2 for linear models), ensuring that dimensionality choices were validated empirically rather than tuned to the training data.

Finally, because interpretability was essential for policy relevance, the study deliberately avoided opaque dimensionality-reduction transforms (e.g., principal components) in the main analyses. Instead, it favored sparse or shrunken coefficient paths (Lasso/Elastic Net) and feature-importance profiles (random forest/XGBoost) that could be mapped back to concrete school-level levers. This approach respected the constraints of a non-experimental design, no causal claims, careful control of leakage and redundancy, while reducing hyperdimensionality in a manner consistent with transparent, reproducible educational analytics.

Population and Sample

The study population included all public high schools in Illinois that reported dropout rates to the ISBE during the years 2017, 2018, 2019, 2020, 2021, and 2022. This dataset included aggregate information about dropout rates and various related factors for each high school rather than individual student data. This school-level data included variables such as dropout rates, student demographics, socioeconomic indicators, school performance metrics, and contextual factors. Analyzing this data helped identify trends and factors influencing dropout rates at the school level during the COVID-19 pandemic (ISBE, 2023).

This study used existing data from the ISBE (https://www.isbe.net/ilreportcarddata), which included information about high school dropout rates during COVID-19. The dataset used in this study is publicly available from the Illinois State Board of Education (ISBE). Since the data is made available to the public by the Illinois State Board of Education, no special licensing or permissions are required for its use. The dataset for the study was supplied by the ISBE and comprised data for all public schools in the state of Illinois (K–12). Specifically, the number of schools included in the dataset was as follows: 3,796 in 2017, 3,889 in 2018, 3,859 in 2019, 3,864 in 2020, 3,856 in 2021, and 3,846 in 2022. The number of schools reported for 2017 through 2019 were categorized by grade levels and school types (i.e., pre-K, elementary, middle, and high school). The same classification applied to 2020 through 2022.

In 2017, there were 644 high schools, 604 middle schools, 2,406 elementary schools, 142 charter schools, and 0 pre-K schools. In 2018, there were 653 high schools, 603 middle schools, 2,398 elementary schools, 141 charter schools, and 93 pre-K schools. In 2019, there were 650 high schools, 602 middle schools, 2,383 elementary schools, 141 charter schools, and 82 pre-K schools (see Table 1).

In 2020, there were 651 high schools, 601 middle schools, 2,386 elementary schools, 138 charter schools, and 87 pre-K schools. In 2021, there were 648 high schools, 603 middle schools, 2,384 elementary schools, 135 charter schools, and 85 pre-K schools. In 2022, there were 646 high schools, 603 middle schools, 2,378 elementary schools, 134 charter schools, and 84 pre-K schools (see Table 2).

In 2017, there were 1,968,196 students with an average of approximately 519 per school. In 2018, there were 5,971,229 students with an average of approximately 1,257 students per school. In 2019, there were 5,909,176 students with an average of approximately 1,248 students per school. In 2020, there were 3,804,807 students with an average of approximately 1,229 students per school. In 2021, there were 3,393,069 students with an average of approximately 1,185 students per school. In 2022, there were 5,545,306 students with an average of approximately 1,179 students per school (see Table 4).

Data Scope and Unit of Analysis

The analysis used school-level administrative records; each observation represented a public high school in Illinois for a given year (2017–2022), and the dependent variable was the annual dropout rate, an aggregate proportion derived from individual student outcomes. Predictors likewise summarized school characteristics (e.g., mobility rate, proportions of low-income or IEP students, attendance/truancy metrics, and contextual indicators such as the COVID-19 period). Because the unit of analysis was the school–year, all estimates and predictions were interpreted at that same level. The models characterized how differences in school-level conditions were associated with differences in school-level dropout rates. They did not estimate the probability that a particular student would drop out, nor did they support individual-level causal claims. This distinction helped avoid ecological fallacy (i.e., inferring

individual behavior from aggregates) and clarified the appropriate use of results for institutional planning and resource allocation, rather than for student-specific decisions.

A practical implication is that the models reflected the average relationship across schools. Large and small schools contributed one observation each, so findings should be interpreted as school-level effects, not enrollment-weighted, student-level effects. Moreover, the bounded and sometimes zero-inflated nature of the outcome at the school level (especially in small schools) introduced additional variability. This variability was handled through robust error metrics and ensemble models but still constrained the granularity of inference. Future work with student-level longitudinal data (e.g., logistic or survival models within a multilevel framework) could bridge the levels by jointly modeling individual risk and school context.

Modeling Suite and Implementation

This study adopted a suite of supervised learning models to predict the annual high school dropout rate from demographic, geospatial, and school-level factors, balancing transparent linear approaches with flexible tree-based ensembles. The modeling work followed CRISP-DM and used a common pipeline so that results were comparable across methods. After imputation and encoding, numeric predictors were standardized when required, and all transformations were fit only on the training folds to avoid leakage. The dataset was partitioned into an 80% training set and a 20% test set, and 10-fold cross-validation on the training data was used to tune hyperparameters. Model selection relied on held-out performance using R^2, *MAE*, and *RMSE*, and model outputs included fitted objects, test-set predictions in original units, residual diagnostics, and either coefficient tables or feature-importance profiles for interpretation.

Multiple linear regression (MLR) served as a baseline for interpretability, modeling a linear relationship between predictors and the dropout rate. It yielded directly interpretable coefficients that quantified the expected change in the outcome for a unit change in each predictor, holding others constant. Although fast and communication friendly, MLR is sensitive to multicollinearity, outliers, and misspecification, and it does not discover nonlinearities or higher-order interactions without explicit feature engineering (J. Y. Chan et al., 2021). In this study, predictors for MLR were standardized, the model was trained within cross-validation, and final coefficients, predictions, and residual plots were produced for the test set.

To stabilize linear estimates in the presence of correlated school-level predictors, Ridge regression augmented MLR with an *L2* penalty that shrank coefficients toward zero. This reduced variance and typically improved generalization when multicollinearity was present, while retaining all variables in the model. Ridge remained linear, and thus interpretable, but introduced bias through shrinkage. Its penalty strength (α) was selected via cross-validated grid search before refitting on the full training set and evaluating on the hold-out test set.

Lasso regression applied an *L1* penalty that could shrink some coefficients exactly to zero, performing embedded feature selection. This yielded more parsimonious models and sharpened substantive conclusions about which factors mattered most. However, when predictors were highly collinear, Lasso sometimes selected one among several near-equivalent variables somewhat arbitrarily. Like Ridge, it preserved the linear form. The penalty parameter was tuned with cross-validation, after which the final model was refit on the training data and assessed on the test set with the same metrics.

Elastic Net combined *L1* and *L2* penalties to balance Lasso's sparsity with Ridge's stability, which was advantageous when predictors formed correlated groups (as was common

with related school and demographic indicators). Two hyperparameters, the overall penalty and the mixing ratio, were tuned within cross-validation. The selected model was refit and evaluated identically to the other linear methods. In practice, Elastic Net often outperformed plain MLR and could match or exceed Ridge/Lasso when collinearity was substantial, while remaining interpretable through shrunken coefficients and selected features (J. Y. Chan et al., 2021).

Random forest represented the non-parametric ensemble family, building many decision trees on bootstrap samples and selecting a random subset of predictors at each split. This structure captured nonlinearities and interactions without manual specification, was relatively robust to outliers, and accommodated mixed data types (Talatahari et al., 2025). Its drawbacks included lower global interpretability, a larger computational footprint, and the absence of simple coefficients (Al-Fakih et al., 2024). In this study, key hyperparameters, such as the number of trees, maximum depth, and node-size constraints were tuned by cross-validation, and the final forest was evaluated on the hold-out set. Feature importance was reported using impurity- or permutation-based measures to aid interpretation.

XGBoost built trees sequentially, with each tree correcting residual errors from prior trees, while regularization and subsampling controlled complexity and prevented overfitting. It often achieved strong accuracy on structured tabular data, handled missing values natively at split time, and provided rich importance diagnostics (Nicho et al., 2025). However, it was more sensitive to hyperparameters and required post hoc tools for interpretability (Punyangarm & Chotayakul, 2025). Tuning focused on the learning rate, number of trees, maximum depth, and sampling rates, which were selected via cross-validation before final training and test-set evaluation. Outputs included test predictions, residual analyses, and gain- or permutation-based importance summaries that identified the most influential predictors.

Across models, inputs included the curated feature matrix and the continuous dropout rate, with preprocessing applied consistently inside cross-validated pipelines to avoid leakage. Outputs were delivered in original units and included comparative metrics (R^2, *MAE*, *RMSE*), error diagnostics, and either coefficients (for the linear family) or variable-importance profiles (for the ensembles). Linear models prioritized transparency and parsimony, Ridge for stability under collinearity, Lasso for embedded selection, and Elastic Net for grouped predictors, whereas random forest and XGBoost prioritized predictive accuracy by discovering nonlinearities and interactions. Final model choice was determined by held-out performance, stability across folds, and interpretability needs, ensuring that conclusions were both empirically sound and practically communicable.

Evaluation Metrics

After training the models, evaluation proceeded along two complementary traditions. In the traditional statistical framing of multiple linear regression, emphasis was placed on goodness of fit and parsimony, chiefly R^2, together with residual diagnostics and assumption checks. In the machine learning framing, emphasis shifted to out-of-sample predictive accuracy, reported through *MAE*, *MSE*, *RMSE*, and, where appropriate, *MAPE*, all computed on cross-validated and held-out test predictions to avoid optimistic bias.

R^2 (coefficient of determination) quantified the proportion of variance in the dropout rate explained by the predictors. An R^2 closer to 1 indicated a better fit, whereas values near 0 (or negative on the test set) indicated limited explanatory power. Because R^2 tends to increase with the addition of predictors, it was interpreted primarily on held-out data rather than on the training set. Adjusted R^2 extended this idea by penalizing superfluous predictors and thus provided a more conservative measure of fit for the linear models. Adjusted R^2 was particularly informative

when comparing variants of multiple linear regression (e.g., Ridge, Lasso, Elastic Net), where parsimony mattered alongside fit.

When prediction accuracy was the priority, the study relied on error-based metrics. Mean absolute error (*MAE*) reported the average magnitude of errors in the original units of the dropout rate, offering an easily interpretable measure, smaller values indicated more accurate predictions. *MAE* is relatively robust to outliers and therefore provided a stable summary of typical error (Iida, 2024). Mean squared error *(MSE)* averaged squared residuals; by squaring errors, it placed extra weight on larger mistakes. The square root of *MSE*, root mean squared error (*RMSE*), returned that penalty to the outcome's original scale and was more sensitive than *MAE* to large deviations. In practice, *MAE* summarized typical performance, while *RMSE* highlighted whether the models made occasional large errors. Consistent improvements in both measures indicated broad predictive gains, not just reductions in a few extreme misses.

When scale-free interpretation was desirable, the study considered mean absolute percentage error (*MAPE*), which expresses errors as percentages of the observed values. Because *MAPE* is undefined (or unstable) when the actual value is zero or very close to zero, and because dropout rates could be zero for some schools, it was interpreted cautiously. Where reported, *MAPE* was calculated with safeguards (e.g., adding a small constant to the denominator) and was treated as complementary to *MAE* and *RMSE*, rather than as a primary selection criterion.

These regression metrics were chosen because the study's outcome, the annual high school dropout rate, was continuous. Consequently, classification metrics, such as precision, recall, and F_1 score were not applicable. Those measures require discrete class labels (e.g., "dropout" vs. "no dropout" or "high-risk" vs. "low-risk") and evaluate a model's ability to recover the positive class under a chosen decision threshold. Applying a threshold to a

continuous target would have discarded information about the magnitude of dropout and would have introduced arbitrary cut points unrelated to the modeling objectives. For this reason, the analysis remained within a regression framework and reported regression-appropriate metrics across all models.

High School Dropout Rate Trends

According to the National Center for Education Statistics (2023), over the decades, the high school dropout rate has seen a notable decline. It decreased from 27.2% in 1960 to 15% in 1970, then declined slightly to 14.1% in 1980. By 1990, it had dropped to 12.1%. The high school dropout rate declined significantly from 9.7% in 2006 to 5.1% in 2021 (National Center for Education Statistics, 2024). Despite this notable decrease, high school dropout remains a persistent problem, as dropouts represent a substantial population (Institute of Education Sciences, 2021).

Table 1

Number of Schools per Level of Education in Report Card Public Dataset for Years 2017 Through 2019

Year	High School	Middle School	Elementary School	Charter School	Pre-K School
2017	644	602	2383	141	82
2018	653	603	2398	141	93
2019	650	604	2406	142	0

Table 2

Number of Schools Per Level of Education in Report Card Public Dataset for years 2020 through 2022

Year	High School	Middle School	Elementary School	Charter School	Pre-K School
2020	651	601	2,386	138	87
2021	648	603	2,384	135	85
2022	646	603	2,378	134	84

The entire census data, rather than a representative sample, was used to conduct this study. Using a census resulted in a more comprehensive representation of the population under study, considering that the research tools supported the analysis of large datasets. Crosnoe (2009) demonstrated the usefulness and practicality of this approach when using secondary data for a group of individuals whose data is difficult to collect individually.

The reduced report card public dataset from the ISBE was obtained with data for the years 2017, 2018, 2019, 2020, 2021, and 2022 for analysis. In 2017, there were 644 high schools and 34 variables. In 2018, there were 653 high schools and 34 variables. In 2019, there were 650 high schools and 34 variables. In 2020, there were 651 high schools and 34 variables. In 2021, there were 648 high schools and 34 variables. In 2022, there were 646 high schools and 34 variables (see Table 3).

Table 3

Number of High Schools and Variables in the Reduced Report Card Public Dataset

Year	Number of High Schools	Number of Variables
2017	644	34
2018	653	34
2019	650	34
2020	651	34
2021	648	34
2022	646	34

Table 4

Number of Students Each Year from 2017 to 2022

Year	Total Number of Students	Average Number of Students per School
2017	1,968,196	519
2018	5,971,229	1,257
2019	5,909,176	1,248
2020	3,804,807	1,229
2021	3,393,069	1,185
2022	5,545,306	1,179

Instrumentation

The report card public dataset used to conduct this study was already preprocessed to exclude personally identifiable information (PII) of the students (ISBE, 2023). Specifically, the dataset was prepared to remove students' personal identification data to guarantee data security and privacy. This meant that no individual names, addresses, or identification numbers that could be linked back to individuals were included.

Data Analysis Tools and Platform

A flexible platform supporting various programming languages and data analysis tools was used in this study. The workstation supported ample computational resources to handle large datasets and efficiently perform complex data analysis. The integrated development environment (IDE) used was Jupyter Notebook. This environment is versatile and seamlessly supports multiple programming languages, such as Python and R. Since Jupyter Notebook supports interactive coding, visualization, and documentation within a single environment, it was well suited for this study. RStudio, which offers robust capabilities for code editing, debugging, and integration with numerous extensions, was also used as a secondary IDE. Python and R were

selected because they are freely available open-source platforms with powerful data processing capabilities.

To conduct the data analysis and implement the predictive models, several Python libraries were utilized due to their robustness, efficiency, and ease of use. NumPy and Pandas were employed for data manipulation, enabling efficient handling of numerical operations and structured datasets. Matplotlib and Seaborn supported data visualization through the creation of plots, charts, and correlation heatmaps, which were essential for exploratory data analysis and evaluating model outputs. Scikit-learn served as the primary library for building and evaluating the multiple linear regression and random forest regression models, offering tools for model training, cross-validation, and performance assessment. Additionally, the XGBoost library was used to implement the XGBoost regression model, providing optimized gradient boosting techniques and feature importance evaluation. Together, these libraries created a comprehensive and reliable environment for developing, validating, and comparing predictive models related to high school dropout rates.

Operational Definitions of Variables

The variables used in this study included both categorical and continuous variables. These variables were:

- County (*CT*)
- District type (*DT*)
- School type (*ST*)
- District size (*DS*)
- Grades served (*GS*)
- Number of student enrollment (*NSE*)

- Rate of student enrollment White (*RSEW*)
- Rate of student enrollment Black or African American (*REBAA*)
- Rate of student enrollment Hispanic or Latino (*RSEH*)
- Rate of student enrollment Asian (*RSEA*)
- Rate of student enrollment with disabilities (*RSED*)
- Rate of student enrollment English learners (*RSEEL*)
- Rate of student enrollment with low income (*RSELI*)
- Rate of student enrollment homeless (*RSEH*)
- Student attendance rate (*SAR*)
- Student mobility rate (*SMR*)
- Student mobility rate English learners (*SMREL*)
- Student mobility rate with IEP (*SMRIEP*)
- Chronically truant students (*CTS*)
- Student chronic truancy rate (*SCTR*)
- High school dropout rate total (*HSDRT*)
- High school dropout rate male (*HSDRM*)
- High school dropout rate female (*HSDRF*)
- High school dropout rate White (*HSDRW*)
- High school dropout rate Black or African American (*HSDRBAA*)
- High school dropout rate Hispanic or Latino (*HSDRHL*)
- High school dropout rate Asian (*HSDRA*)
- High school dropout rate English learners (*HSDREL*)
- High school dropout rate with IEP (*HSDIEP*)

- High school dropout rate with low income (*HSDRLI*)
- Average class size high school (*ACSHS*)
- Average number of days of physical education per week per student (*ANDPEWS*)
- Teacher retention rate (*TRR*)
- Principal turnover within six years (*PTWSY*)

Dataset Fields

The fields in the dataset were operationally defined as follows:

County (*CT*): One of the independent variables in this study. *CT* refers to an administrative division in the state of Illinois and is a nominal variable.

District Type (*DT*): One of the independent variables in this study. *DT* refers to the type of school district in Illinois. A school district is organized and established to provide instruction up to and including grade 12 (ISBE, 2023). *DT* is a nominal variable.

School Type (*ST*): One of the independent variables in this study. *ST* refers to the type of school within a district and is a nominal variable.

District Size (*DS*): One of the independent variables in this study. *DS* refers to the size (small, medium, or large) of a school district in Illinois (ISBE, 2023). *DS* is an ordinal variable.

Grades Served (*GS*): One of the independent variables in this study. *GS* refers to the grade levels provided by a school and/or district for general education (ISBE, 2023). This study focused on grades 9 through 12. *GS* is a nominal variable.

Number of Student Enrollment (*NSE*): One of the independent variables in this study. *NSE* refers to the total student enrollment in the school and district as of October 1 of the school year (ISBE, 2023). *NSE* is a ratio variable.

Rate of Student Enrollment White (*RSEW*): One of the independent variables in this study. *RSEW* refers to the proportion of White students enrolled in the school and district as of October 1. *RSEW* is a ratio variable.

Rate of Student Enrollment Black or African American (*REBAA*): One of the independent variables in this study. *REBAA* refers to the proportion of Black or African American students enrolled in the school and district as of October 1 (ISBE, 2023). *REBAA* is a ratio variable.

Rate of Student Enrollment Hispanic or Latino (*RSEHL*): One of the independent variables in this study. *RSEHL* refers to the proportion of Hispanic or Latino students enrolled in the school and district as of October 1 (ISBE, 2023). *RSEHL* is a ratio variable.

Rate of Student Enrollment Asian (*RSEA*): One of the independent variables in this study. *RSEA* refers to the proportion of Asian students enrolled in the school and district as of October 1 (ISBE, 2023). *RSEA* is a ratio variable.

Rate of Student Enrollment with Disabilities (*RSED*): One of the independent variables in this study. *RSED* refers to the proportion of students with disabilities enrolled in the school and district as of October 1 (ISBE, 2023). *RSED* is a ratio variable.

Rate of Student Enrollment English Learners (*RSEEL*): One of the independent variables in this study. *RSEEL* refers to the proportion of English learners enrolled in the school and district as of October 1 (ISBE, 2023). *RSEEL* is a ratio variable.

Rate of Student Enrollment with Low Income (*RSELI*): One of the independent variables in this study. *RSELI* refers to the proportion of students from low-income families enrolled in the school and district as of October 1 (ISBE, 2023). *RSELI* is a ratio variable.

Rate of Student Enrollment Homeless (*RSEH*): One of the independent variables in this study. *RSEH* refers to the proportion of homeless students enrolled in the school and district as of October 1 (ISBE, 2023). *RSEH* is a ratio variable.

Student Attendance Rate (*SAR*): One of the independent variables in this study. *SAR* refers to the weighted measure of the number of days a student is present relative to the total number of potential attendance days (ISBE, 2023). Poor attendance can lead to high school dropout (van Den Berghe et al., 2022). Roure et al. (2021) stated that students with good attendance who show interest in their academic work increase their chances of success. *SAR* is a ratio variable.

Student Mobility Rate (*SMR*): One of the independent variables in this study. *SMR* refers to the unduplicated count of students who transferred in and out of the serving school at any time during the school year (ISBE, 2023). Higher student mobility can lead to lower performance and increased dropout risk (Lickteig & Lickteig, 2019). *SMR* is a ratio variable.

Student Mobility Rate English Learners (*SMREL*): One of the independent variables in this study. *SMREL* refers to the proportion of English learner students who transferred in and out of the serving school during the academic year (ISBE, 2023). *SMREL* is a ratio variable.

Student Mobility Rate with IEP (*SMRIEP*): One of the independent variables in this study. *SMRIEP* refers to the proportion of students with IEPs who transferred in and out of the serving school during the academic year (ISBE, 2023). *SMRIEP* is a ratio variable.

Chronically Truant Students (*CTS*): One of the independent variables in this study. *CTS* refers to the number of students subject to compulsory attendance who were absent without valid cause for 5% or more of the previous 180 regular attendance days (ISBE, 2023). *CTS* is a ratio variable.

Student Chronic Truancy Rate (*SCTR*): One of the independent variables in this study. *SCTR* refers to the proportion of students subject to compulsory attendance who were absent without valid cause for 5% or more of the previous 180 regular attendance days (ISBE, 2023). *SCTR* is a ratio variable.

High School Dropout Rate Total (*HSDRT*): The response variable in this study. According to Lee-St. John et al. (2018), students are considered dropouts when they fail to complete high school. Dropouts are students who are of school age but are no longer enrolled or did not graduate from high school (ISBE, 2023). McCauley (2017) stated that the rate of students with IEPs who leave high school is higher than that of their regular peers. According to the ISBE (2023), *HSDRT* refers to the proportion of high school dropouts in an entity (school, district, state) per enrollment. *HSDRT* is a ratio variable.

High School Dropout Rate Male (*HSDRM*): One of the independent variables in this study. *HSDRM* refers to the proportion of male students who are of school age but are no longer enrolled or did not graduate from high school (ISBE, 2023). *HSDRM* is a ratio variable.

High School Dropout Rate Female (*HSDRF*): One of the independent variables in this study. *HSDRF* refers to the proportion of female students who are of school age but are no longer enrolled or did not graduate from high school (ISBE, 2023). *HSDRF* is a ratio variable.

High School Dropout Rate White (*HSDRW*): One of the independent variables in this study. *HSDRW* refers to the proportion of White students who are of school age but are no longer enrolled or did not graduate from high school (ISBE, 2023). *HSDRW* is a ratio variable.

High School Dropout Rate Black or African American (*HSDRBAA*): One of the independent variables in this study. *HSDRBAA* refers to the proportion of Black or African

American students who are of school age but are no longer enrolled or did not graduate from high school (ISBE, 2023). *HSDRBAA* is a ratio variable.

High School Dropout Rate Hispanic or Latino (*HSDRHL*): One of the independent variables in this study. *HSDRHL* refers to the proportion of Hispanic or Latino students who are of school age but are no longer enrolled or did not graduate from high school (ISBE, 2023). *HSDRHL* is a ratio variable.

High School Dropout Rate Asian (*HSDRA*): One of the independent variables in this study. *HSDRA* refers to the proportion of Asian students who are of school age but are no longer enrolled or did not graduate from high school (ISBE, 2023). *HSDRA* is a ratio variable.

High School Dropout Rate English Learners (*HSDREL*): One of the independent variables in this study. *HSDREL* refers to the proportion of English learner students who are of school age but are no longer enrolled or did not graduate from high school (ISBE, 2023). *HSDREL* is a ratio variable.

High School Dropout with IEP (*HSDIEP*): One of the independent variables in this study. *HSDIEP* refers to the proportion of students with IEPs who are of school age but are no longer enrolled or did not graduate from high school (ISBE, 2023). *HSDIEP* is a ratio variable.

High School Dropout Rate with Low Income (*HSDRLI*): One of the independent variables in this study. *HSDRLI* refers to the proportion of students from low-income families who are of school age but are no longer enrolled or did not graduate from high school (ISBE, 2023). Students from low-income families often drop out due to financial difficulties (Goga et al., 2021). *HSDRLI* is a ratio variable.

Average Class Size High School (*ACSHS*): One of the independent variables in this study. *ACSHS* refers to the average number of students in each high school class as of the last day of school (ISBE, 2023). *ACSHS* is a ratio variable.

Average Number of Days of Physical Education per Week per Student (*ANDPEWS*): One of the independent variables in this study. *ANDPEWS* refers to the average number of days per week a student engages in physical education (ISBE, 2023). *ANDPEWS* is a ratio variable.

Teacher Retention Rate (*TRR*): One of the independent variables in this study. *TRR* refers to the proportion of full-time teachers who remained in the same school or district over the past three years (ISBE, 2023). *TRR* is a ratio variable.

Principal Turnover Within Six Years (*PTWSY*): One of the independent variables in this study. *PTWSY* refers to the number of different principals who served at the same high school within the past 6 years (ISBE, 2023). *PTWSY* is a ratio variable.

Confounding variables, such as parental involvement, socioeconomic factors, and teacher quality, were identified during the literature review (Lorenzo-Lledó et al., 2020). For instance, Koç et al. (2020) demonstrated that school dropout factors included adjustment problems at the individual, gender, family, teacher, economic, and macrosystem levels. Similarly, Pov et al. (2020) suggested that grade retention, absenteeism, academic expectations, and engagement in private tutoring had substantial impacts on dropout rates. To account for these factors, this study utilized several variables in the dataset classified as confounding, as described under the operational definitions of variables. To mitigate confounding variables, school-level covariates were included as predictors in all models, and multicollinearity was addressed via correlation and VIF screening. Model influence was assessed through cross-validated feature importance and residual diagnostics.

Study Procedures

The study procedure included data collection, data preprocessing, exploratory data analysis, feature selection, and modeling (i.e., model training, model evaluation, and model comparison).

Data Collection

The 2017, 2018, 2019, 2020, 2021, and 2022 report card public datasets from the ISBE, which include information about high school dropout rates before and during COVID-19, were obtained from the ISBE website (https://www.isbe.net/ilreportcarddata). These datasets are public and open-access.

Data Preprocessing and Feature Engineering

The raw files for 2017–2022 were merged into a single dataset, column names were standardized (trimmed and lowercased), and string fields between *county* and *grades_served* were cleaned for stray whitespace and casing. Records were restricted to high schools by filtering school_type == "High School", and indices were reset. Two derived fields were prepared for modeling: *district_size_encoded* (mapping Small/Medium/Large to 1/2/3) and *covid_period* (encoding calendar year into a binary indicator: pre-Covid (2017-2019) vs. during Covid (2020–2022)).

A missing data audit combined counts, percentages, and a heatmap to guide feature retention. In accordance with the study plan, variables with more than 50% missingness were removed prior to modeling. The excluded set included *grades_served* and several subgroup-specific indicators with sparse coverage across years (e.g., *mobility_rate_el*, *mobility_rate_iep*, and subgroup dropout rates, such as *dropout_rate_white*, *dropout_rate_black*,

dropout_rate_hispanic, *dropout_rate_female*, *dropout_rate_male*, *dropout_rate_iep*, *dropout_rate_low_income*, *dropout_rate_asian*), as well as subgroup enrollment counts (e.g., *enrolment_el*, *enrolment_asian*, *enrolment_black*, *enrolment_hispanic*, *enrolment_white*). These variables were excluded due to high missingness, which posed a risk of unstable imputation and biased estimates. Additionally, many subgroup dropout rates were highly redundant with the overall outcome.

Beyond missingness, the dataset was screened for multicollinearity using the correlation matrix and variance inflation factor (VIF). Two predictors were dropped on this basis. First, *year* was removed due to strong overlap with *covid_period* (consistent with its binary definition) and a VIF of approximately 7.73, which indicated non-trivial redundancy. Retaining only *covid_period* simplified the interpretation of pre- vs. during-COVID contrasts. Second, *chronically_truant* was removed due to its high collinearity with *chronic_truancy_rate*. All remaining predictors showed acceptable pairwise correlations and VIF values and were retained.

Imputation followed a simple, robust scheme aligned with the data structure and modeling goals. For numeric predictors, missing values were replaced with the median of each variable (computed from available data), which mitigated skew and outlier influence while preserving scale. This imputation was variable-specific and designed for compatibility across linear and tree-based models. Numeric variables with missing data included *enrolment_iep* (51), *enrolment_low_income* (4), *attendance_rate* (31), *mobility_rate* (1), *dropout_rate_total* (2), *class_size_high* (33), *pe_days* (128), *retention_rate* (19), and *turnover* (17). These were imputed with their respective medians, computed on training folds and applied to validation and test sets to avoid leakage. For the encoded categorical variable *district_size_encoded*, 36 missing values were imputed using the mode (i.e., most frequent category).

Although tree-based imputers (e.g., random forest–style iterative imputation) can capture nonlinear dependencies, the median/mode strategy was preferred for reproducibility and simplicity, given that Random Forest and XGBoost are inherently robust to distributional irregularities. Sensitivity checks may be added in an appendix if a more complex imputer is adopted later.

Potential outliers were examined using standardized scores ($|Z| > 3$) on numeric fields. No automated trimming or winsorization was applied. Instead, robustness was addressed through model choice (ensemble methods) and through diagnostic checks (e.g., residual distributions, residuals-versus-fits plots) to ensure that extreme observations did not unduly influence results. Transformations used for exploratory visualization (e.g., standardization and Yeo–Johnson transformation for skew) were documented separately and not used during modeling to avoid data leakage. The modeling data used for training, validation, and testing preserved the outcome variable in its original units to ensure that *MAE*, *RMSE*, and related metrics remained interpretable at the school-year level.

Exploratory Data Analysis

Exploratory data analysis helped the researcher identify missing values, outliers, and incomplete records (Ayodele, 2023). This process enabled a preliminary investigation of the dataset and summarized its main characteristics (Avval et al., 2021). Measures of central tendency were computed to describe the distribution's central location, and measures of dispersion were used to examine variability. Skewness and kurtosis were computed to evaluate the shape of the distribution. Histograms and boxplots were generated to visualize these distributions (Newburger et al., 2023). Calculating these metrics and creating the plots helped determine whether transformations were necessary before data analysis.

Data Cleaning

Missing data were handled in two steps: variables with more than 50% missingness were removed, and remaining gaps were imputed using the median for numeric variables and the mode for the encoded categorical variable. However, during the data cleaning process, several variables were found to have more than 50% missing data, rendering them unsuitable for reliable imputation. As a result, these variables were excluded from the final analysis to maintain the integrity and validity of the dataset. Specifically, the following dropout rate variables were removed: high school dropout rate for Asian students, Native Hawaiian or Other Pacific Islander, American Indian or Alaska Native, and students identifying as two or more races. Although the ISBE mandates dropout reporting from each school district, the prevalence of missing data in these subgroups limited their utility in this study.

Similarly, missing values for variables unrelated to dropout rates could negatively impact the reliability and generalizability of results (Hegde et al., 2019). As a result, these variables were replaced with appropriate figures generated from the random forest method. The predictor variables included demographic, geolocation, and internal school factors. For categorical variables, categorical encoding was applied to convert attributes into numerical values.

The researcher renamed the variables used in the study to facilitate analysis. Categorical variables were assigned numerical values to enable statistical computation. For example, the predictor variable *district size* (DT), which included three categories (small, medium, and large), was encoded as 1, 2, and 3, respectively, to reflect their ordinal relationship. This transformation was conducted to support statistical analysis. The datasets were uploaded into RStudio to perform exploratory data analysis (EDA) (Rahmany et al., 2020).

According to Rahmany et al. (2020), RStudio is effective and user-friendly for manipulating and visualizing data in R. The researcher cleaned the datasets for any missing values and outliers and ensured the assumptions for applying multiple linear regression (MLR) and random forest (RF) were satisfied.

Feature Selection

Three distinct models were used to analyze the variables: multiple linear regression, random forest, and XGBoost regression. Pearson correlation was used to examine linear relationships between individual predictor variables, providing insight into the strength and direction of associations. This analysis helped assess how each predictor independently influenced the high school dropout rate. In contrast, the predictive models, MLR, random forest, and XGBoost, enabled a more comprehensive analysis by incorporating interactions and combined effects. These models predicted the dropout rate using multiple predictor variables simultaneously, capturing complex and nonlinear relationships. Consequently, while linear regression analyses isolated individual variable effects, the predictive models offered a broader perspective on how multiple factors collectively influenced dropout outcomes. The researcher implemented the machine learning algorithms (MLR, XGBoost, and RF), evaluated each, and compared them to determine which model best predicted the response variable.

Modeling

Model Training

Predictor variables included demographic, geolocation, and internal school factors. The response variable was the high school dropout rate in the state of Illinois. For initial model development, an 80/20 train–test split was applied: 80% of the data were used for training and 20% for testing, consistent with recommendations by Gholamy et al. (2018). Additionally, K-

fold cross-validation was implemented for all three models, MLR, RF, and XGBoost, to ensure robust performance evaluation. This method allowed the models to be trained and validated across multiple data subsets, providing more accurate and stable predictive accuracy estimates while reducing the risk of overfitting (Nti et al., 2021).

Evaluation metrics, such as mean absolute error (MAE), *MSE*, root mean squared error (RMSE), R^2, and mean absolute percentage error (MAPE) were used for all models. Adjusted R^2 was reported only for the MLR model, as it is not applicable to non-parametric models, such as RF and XGBoost.

Model Validation

A cross-validation technique was used to validate the models. K-fold cross-validation was implemented to split the dataset into 10 folds or subsets of approximately equal size (Nti et al., 2021). The model was trained on nine folds and tested on the remaining fold. This process was repeated 10 times, with each fold serving as the test set once. The performance from each iteration was averaged to obtain an overall estimate of model performance (Nti et al., 2021).

Model Evaluation

After validating the MLR, RF, and XGBoost models, the next step involved evaluating their predictive performance using appropriate regression metrics. To provide a comprehensive assessment of each model, the following evaluation criteria were applied: MAE, MSE, RMSE, R^2, and MAPE. Additionally, adjusted R^2 was reported for the MLR model, as it accounts for the number of predictors and is specific to parametric models. Feature importance rankings were extracted from the RF model to identify the most influential predictors. These metrics enabled a robust comparison of each model's accuracy, generalizability, and practical utility.

Data Modeling

Using multiple linear regression to analyze data enables clear interpretation. This model is flexible and adaptable. However, linear regression is sensitive to outliers and noise and is prone to overfitting and underfitting. While multiple regression models help analyze the influence of predictor variables on the response variable, improperly applied analysis of large datasets can lead to false conclusions (Liu et al., 2022).

Using random forest helps build a forest with an ensemble of decision trees. This model is an easy-to-use machine learning algorithm that reduces overfitting in decision trees and improves accuracy. The flexibility of this model to handle both classification and regression problems makes it suitable for a broader range of applications. Additionally, the random forest model works well with both categorical and continuous variables and can automate the handling of missing data. However, it does not provide complete visibility into coefficients as linear regression does. Random forest is also computationally intensive, especially when applied to large datasets (Langsetmo et al., 2023).

Using the XGBoost regression model provides a highly accurate and efficient approach for predicting outcomes in structured data. This model is an advanced ensemble learning technique that builds upon gradient boosting principles, whereby trees are added sequentially to correct the errors of prior trees. XGBoost is known for its ability to model both linear and nonlinear relationships, its built-in regularization to prevent overfitting, and its strong predictive performance on large and complex datasets. It supports both categorical and continuous variables and handles missing data internally. However, like random forest, XGBoost is often considered a “black box” model, offering limited interpretability compared to linear regression. It also

requires careful hyperparameter tuning and can be computationally intensive, particularly when applied to very large datasets or when high precision is desired.

Traditional methods used by school districts to examine and mitigate school dropout have not been as effective as modern techniques. These methods typically focused on student-level variables, failing to account for interactions between students and their environments (Allensworth & Easton, 2007; Balfanz et al., 2007; Neild, 2009). For instance, Early Warning Systems have produced some reliable results in identifying at-risk students; however, they often rely on data that fail to capture the full range of contributing factors.

Machine learning algorithms use data to identify at-risk students early so that dropout may be prevented. These algorithms can predict student dropout with accuracy by harnessing large datasets and advanced analytical techniques. Machine learning models have demonstrated the ability to accurately identify students who are more likely to drop out by examining diverse factors and detecting complex patterns (Song et al., 2023). These models have significantly advanced the prevention of school dropout and helped address its economic and social impacts (Segura et al., 2022).

In this study, multiple linear regression, a machine learning algorithm used for supervised learning, was used to predict high school dropout rates based on demographic, geolocation, and other internal school factors. This statistical method can analyze the relationship between a single response variable and multiple predictor variables. It identifies which independent variables significantly influence the response variable and predicts the value of the dependent variable using known predictors (Adasme et al., 2023). This model also helps detect outliers and anomalies and highlights which predictors are strongly correlated with the outcome.

Despite the strengths of this model, several drawbacks exist. These include susceptibility to outliers, limited flexibility, overfitting, multicollinearity assumptions, and the assumption of linearity. The model assumes a linear relationship between predictors and the outcome, which may not hold when the relationship is nonlinear, in which case more complex models may be needed.

The random forest regression model, a supervised machine learning algorithm, was also used to predict high school dropout rates based on demographic, geolocation, and internal school variables. This model, an ensemble learning technique, is especially useful for classification and regression problems. It is robust to outliers and multicollinearity and addresses overfitting and high variance commonly found in decision trees (Jimenez et al., 2023). It can handle various data types, aggregates results across multiple decision trees and generally outperforms many other learning algorithms (Dass et al., 2021).

Despite its strengths, the random forest model also has limitations. These include reduced transparency in describing data relationships, high computational complexity, memory usage, and the risk of overfitting in large ensembles. Random forest may also become slow or ineffective in prediction when a large number of trees are used, and it requires significant memory when applied to large datasets.

The XGBoost regression model, also a supervised machine learning algorithm, was included in this study to predict high school dropout rates based on demographic, geolocation, and internal school variables. XGBoost is an advanced ensemble method that builds trees sequentially, with each new tree correcting residual errors from the previous one. It incorporates regularization, parallel processing, and automatic handling of missing data, making it efficient and scalable for large, complex datasets. XGBoost is particularly effective at modeling nonlinear

relationships and capturing intricate interactions among predictors, thereby improving predictive accuracy. It also provides mechanisms for evaluating feature importance, offering insight into which factors most influence dropout rates.

Despite the numerous advantages of the XGBoost model, several limitations exist in its implementation. These include its black-box nature, which makes it difficult to interpret the internal mechanics of the model compared to more transparent models like linear regression. In addition, the model's predictive strength is highly dependent on hyperparameter tuning, which can be time-consuming and computationally demanding. While XGBoost can outperform many traditional models in terms of accuracy, it requires significant memory and processing power when applied to high-dimensional data, especially during extensive training cycles.

This study aimed to predict the high school dropout rate based on demographic, geolocation, and other internal school factors. Choosing multiple linear regression, random forest, and XGBoost to conduct this study was appropriate, as all three models can predict a continuous response variable from multiple predictors. The dataset used contained both categorical and continuous values. Multiple linear regression is flexible, adaptable, and provides easily interpretable outputs. Random forest works well with both categorical and continuous values and can handle missing data. Similarly, the XGBoost model offers high predictive performance, efficiently manages missing values, and captures nonlinear relationships and interactions among variables, although it requires more computational resources and careful hyperparameter tuning.

Multiple Linear Regression

Multiple linear regression (MLR) was used to analyze the influence of predictor variables on the continuous response variable. This machine learning algorithm was suitable for this

quantitative study, as it applies when a single continuous outcome is modeled using multiple independent variables (Zhao & Li, 2022). The primary algorithm used was Ordinary Least Squares (OLS) regression. According to Farbo et al. (2024), the conventional method for assessing the relationships between predictor variables (e.g., internal school factors, demographics, geolocation) and the dependent variable (high school dropout rate) is OLS. This method minimizes the sum of squared differences between observed and predicted values to yield comprehensible coefficients that quantify the expected change in the dropout rate for each predictor. When core assumptions, such as linearity, independence, homoscedasticity, and residual normality are satisfied, OLS is a widely accepted and effective statistical method.

The assumptions for multiple regression, linearity, homoscedasticity, independence of errors, normality, and non-collinearity among independent variables, were assessed prior to analysis. Diagnostic plots were used to detect violations of these assumptions. If an assumption was not met, appropriate transformations were applied before model fitting.

Since the MLR model in this study was implemented within a machine learning framework, the emphasis shifted from traditional assumption testing to predictive accuracy on unseen data. Evaluation of the ML-based MLR model was conducted using performance metrics commonly used in predictive modeling: R-squared (R^2), adjusted R-squared, *MSE*, root mean squared error (*RMSE*), mean absolute error (*MAE*), and mean absolute percentage error (*MAPE*). Diagnostic plots such as residual plots and predicted versus actual plots were also used to assess model fit and error distribution visually.

The coefficient of determination (R^2) was calculated to assess the explanatory power of the multiple regression model. This metric indicated the proportion of variance in the dependent variable, the high school dropout rate, which could be explained by the selected independent

variables. The R^2 value was derived using cross-validated predictions to prevent overestimation due to overfitting. A higher R^2 indicated that the model accounted for a greater proportion of variance in dropout rates across schools. A score closer to 1 indicated that a large proportion of the variance was explained by the predictors, while a value near zero suggested limited explanatory power. As Kaneko (2017) noted, R^2 serves as a measure of the model's predictive accuracy.

The adjusted R^2 statistic, a modified version of R^2 that penalizes the addition of unnecessary predictors, was also computed. This metric provided a more accurate measure of model fit when multiple predictors were included (Prins et al., 2023). The *MSE*, defined as the average squared difference between observed and predicted values, was also calculated. Because *MSE* gives more weight to large errors, it is sensitive to outliers. The root mean squared error (*RMSE*), the square root of the *MSE*, was computed to express model error in the original units of the outcome variable. As Gomez-Vazquez et al. (2024) noted, *RMSE* provides an interpretable measure of model fit in real-world terms.

Residual analysis was conducted to identify patterns or trends in the residuals. This analysis helped detect potential model issues, including autocorrelation, heteroscedasticity, and nonlinearity. To test for autocorrelation, the Durbin–Watson test was conducted. Heteroscedasticity was assessed using White's test, and nonlinearity was examined visually through boxplots.

Random Forest Regression Model

A random forest regression model, a supervised machine learning algorithm, was also used to predict high school dropout rates based on demographic, geolocation, and internal school factors. This algorithm helps address the problems of overfitting and high variance in decision

trees (Jimenez et al., 2023). In addition, this model is robust and can handle various types of data. A decision tree model was developed using the high school dropout rate as the response variable and the demographic, geolocation, and internal factors as predictor variables.

The decision tree was built by recursively splitting the training dataset, comprising predictor and response variables, into smaller subsets. Splits were made based on predictor variables to maximize homogeneity in the child nodes regarding the outcome variable. The splitting process continued until the leaves contained the predicted output values. The random forest model then aggregated the output from many decision trees to produce the final prediction. This process is known as ensemble learning (Dass et al., 2021).

Random forest was chosen because it can combine the results of multiple decision trees that segment data by demographic, geolocation, and internal school factors to predict dropout rates. Xie et al. (2024) developed and validated an MRI radiomics-based decision support tool using several machine learning models. They found that random forest outperformed other algorithms ($p < .01$). The decision to use random forest in this study was further supported by its ability to handle complex nonlinear relationships and avoid overfitting (Nti et al., 2021).
The following steps were used:

1. The dataset was split into two parts: 80% for training and 20% for testing. The training set was used to build the decision tree, and the testing set was used to evaluate its performance (Bichri et al., 2024; Vrigazova, 2021).
2. K-fold cross-validation (with $k = 10$) was used to validate the model and tune hyperparameters such as the number of trees (N) and maximum tree depth. This technique ensured that the model was trained and validated on multiple data partitions, reducing overfitting and enhancing generalizability.

3. For each training fold, a set of decision trees was built using bootstrapped samples and a random subset of predictors at each split.
4. The number of decision trees (*N*) was predefined (e.g., 100 or more), and tree construction, including random sampling and feature selection, was repeated for each tree in the ensemble.
5. Each tree generated a predicted dropout rate for each test observation.
6. The final prediction was the average of all individual tree predictions, reducing variance and improving accuracy.

A random forest is a group of decision trees whose outputs are combined into a single result. Given that decision trees are supervised learning algorithms, the model was trained using a preprocessed subset of the data. Specifically, 80% of the cleaned dataset was used for training, while the remaining 20% was reserved for testing and performance evaluation (Azad et al., 2022; Sameer & Sriramya, 2021).

The decision tree algorithm is a simple and efficient method in supervised learning, where data points are repeatedly split based on preset parameters related to the research problem (Azad et al., 2022; Sameer & Sriramya, 2021; Ramani et al., 2022). Because decision trees are sensitive to noise, it was essential to remove noisy data before training (Kumar et al., 2016).

After data collection, the dataset was preprocessed and split into training and testing subsets. Eighty percent of the data were used for training, and 20% were used for testing (Ramani et al., 2022). This 80/20 strategy is widely adopted and yields accurate results (Gholamy et al., 2018). Studies investigating the impact of the train/test split ratio on machine learning performance have shown that an 80/20 split yields optimal results (Bichri et al., 2024; Gholamy et al., 2018; Vrigazova, 2021).

XGBoost Regression Model

In this study, an XGBoost regression model, a powerful supervised machine learning algorithm, was used to predict high school dropout rates based on demographic, geolocation, and other internal school factors. XGBoost is an optimized implementation of the gradient boosting framework, which builds additive decision trees sequentially to minimize prediction error (Jinbo et al., 2025). Unlike traditional boosting techniques, XGBoost incorporates regularization to control model complexity and improve generalization, making it well suited for handling structured, high-dimensional data (Tarwidi et al., 2022).

XGBoost was selected for this study due to its ability to model both linear and nonlinear relationships, its robustness to missing values, and its high predictive accuracy. The model automatically handles missing data by learning the best direction to assign missing values at each split, and it accommodates both continuous and categorical variables. Additionally, XGBoost provides feature importance scores, enabling interpretation of which variables contribute most significantly to predicting dropout rates. This is particularly valuable in educational research, where understanding the influence of each factor is as important as achieving accurate predictions (Inoue et al., 2020).

Procedure for XGBoost Implementation

Data Preparation. The dataset was cleaned and preprocessed to resolve inconsistencies. Missing data were handled in two steps: variables with more than 50% missing values were removed, and the remaining gaps were imputed using the median for numeric variables and the mode for the encoded categorical variable. The cleaned dataset was split into 80% training and 20% testing subsets, consistent with best practices in machine learning (Bichri et al., 2024;

Gholamy et al., 2018). Imputers were fit on the training folds only and applied to validation/test data to prevent leakage.

Model Construction. The XGBoost regressor was initialized with baseline hyperparameters. During training, XGBoost sequentially added trees, each minimizing the residual errors of the previous iteration. Hyperparameters such as learning rate, maximum tree depth, subsample ratio, and the number of boosting rounds were tuned using 10-fold cross-validation to improve generalization and reduce overfitting.

Model Evaluation. Once trained, the model was evaluated using the test set. Performance metrics included R^2, adjusted R^2, mean squared error (*MSE*), root mean squared error (*RMSE*), mean absolute error (*MAE*), and mean absolute percentage error (*MAPE*). These metrics ensured consistent comparison with the multiple linear regression and random forest models.

Model Interpretation. Feature importance plots were generated to rank predictors according to their relative contribution to dropout rate prediction. For XGBoost, the study relied on the model's built-in feature importance scores to quantify the influence of each predictor on the trained model's predictions, including the binary *covid_period* indicator (pre-COVID vs. during-COVID). SHAP (SHapley Additive exPlanations) is a widely used approach for attributing prediction influence to individual features; however, SHAP values were not computed for this analysis.

Although XGBoost offers superior performance for many prediction tasks, it has limitations. The model is computationally intensive, especially when working with large datasets or requiring extensive hyperparameter tuning. Additionally, interpretability is limited unless aided by specialized tools. Nevertheless, the model's predictive power, resistance to overfitting,

and ability to capture complex patterns make it a valuable component of this study's modeling strategy (Inoue et al., 2020; Jinbo et al., 2025; Tarwidi et al., 2022).

Model Evaluation

Evaluating decision tree models effectively, beyond splitting data into training, validation, and test sets, requires appropriate performance metrics. For regression tasks, these include *MSE*, *MAE*, and *RMSE* (Abdullah et al., 2024; Aljohani & Aburasain, 2024). In this study, the models were used to predict a continuous variable. After training, the model was fitted to the data, and performance was evaluated. The trees were visualized, and *RMSE* was calculated to assess the goodness-of-fit for both the random forest and XGBoost models (Jalota & Suthar, 2024).

Although metrics such as accuracy, precision, recall, F1 score, and area under the ROC curve (AUC-ROC) are typically used in classification tasks (Kabathova & Drlik, 2021; Melo & de Souza, 2023; Schröer et al., 2021), they are not appropriate for continuous regression models. Therefore, only regression-appropriate metrics were used for model comparison in this study.

After analyzing the data using multiple linear regression, random forest, and XGBoost, a comparison was conducted to determine which model provided the best prediction of high school dropout rates. The McFadden pseudo R^2 was used to compare model accuracy (Scarano et al., 2023). The performance of the three models was measured using metrics including *MAE* and *RMSE* (Jalota & Suthar, 2024). These metrics helped determine which model best predicted the response variable, high school dropout rate (Jalota & Suthar, 2024; Scarano et al., 2023).

Assumptions

Data-Level Assumptions

The researcher assumed that the Illinois Report Card data were consistently defined and measured across school–years from 2017 to 2022 and that the reduced analytic sample was representative of Illinois public high schools over that period. Missingness was assumed to be at least Missing at Random (MAR) with respect to observed covariates. Median (for numeric) and mode (for encoded categorical) imputations were fit on training folds and applied to validation/test data to prevent leakage. Because the outcome, the annual dropout rate, was a bounded proportion with mass at zero for some schools, percentage-based error metrics such as *MAPE* were interpreted cautiously. Emphasis was placed on *MAE* and *RMSE* in original units for interpretability.

At the school–year level, observations were treated as conditionally independent. However, repeated measures for the same schools across years implied potential serial or cluster dependence, which could inflate generalization estimates. To mitigate this, models were evaluated using 10-fold cross-validation on the training set and a held-out test set. A time indicator (*COVID-19 period*) was included to capture structural shifts. Still, unobserved temporal or spatial clustering may have remained.

Multiple Linear Regression (OLS/MLR)

The linear models assumed a correctly specified linear functional form between predictors and the dropout rate, independence of errors across observations, homoscedastic residual variance, and approximately normal residuals for inference. They also assumed no exact multicollinearity among predictors. These assumptions were examined through correlation matrices and variance inflation factor (VIF) screening prior to modeling, followed by residual-

versus-fitted plots, Q–Q plots, and statistical tests such as Durbin–Watson (for autocorrelation) and White or Breusch–Pagan tests (for heteroscedasticity). Because the outcome was bounded, the possibility that OLS could extrapolate outside [0, 100] was monitored in predicted-versus-actual plots. No clipping was applied to predictions, allowing error metrics to reflect true predictive behavior. Anticipated departures from linearity or constant variance supported the inclusion of non-parametric ensemble models.

Ridge, Lasso, and Elastic Net

The regularized linear models shared the core OLS assumptions, linearity in parameters, independent and homoscedastic errors, and approximately normal residuals, while adding penalty terms to stabilize estimation. Ridge assumed many small effects and shrank coefficients toward zero without eliminating predictors. Lasso assumed sparsity and could set some coefficients exactly to zero, thus performing variable selection. Elastic Net combined both approaches to handle groups of correlated predictors. Hyperparameters (penalty strength and, for Elastic Net, the *l1_ratio*) were selected via cross-validated grid search on the training data. Although regularization improved model parsimony and out-of-sample performance, it did not relax the assumptions of linearity or homoscedasticity. When these assumptions were materially violated, tree-based methods were expected to perform better.

Random Forest

The random forest model imposed far fewer parametric assumptions. It assumed that training and test data were drawn from the same distribution and that observations were sufficiently independent for bootstrap resampling and cross-validation to be valid. Random forest did not require linearity or homoscedasticity and was able to accommodate complex interactions and nonlinearities. However, impurity-based importance measures could be biased toward

variables with greater variance or more split points. Thus, variable importance was interpreted diagnostically rather than causally. Model complexity (e.g., number and depth of trees, minimum samples per split/leaf, and feature subsampling) was tuned to reduce overfitting and control variance.

XGBoost

The boosted-tree model similarly assumed distributional stability between training and test sets and conditional independence of observations. It did not assume linearity and explicitly controlled model complexity through learning rate, maximum depth, and row/column subsampling. Missing values were handled natively by learning default split directions, assuming that missingness patterns carried predictive signal without distorting scale. Due to boosting's sensitivity to hyperparameters, cross-validated tuning was used to mitigate overfitting. As with random forest, feature importance rankings were treated as descriptive rather than causal.

Evaluation-Related Assumptions and Checks

All preprocessing steps involving learned parameters (e.g., imputation, scaling) were applied only to training folds and then reused for validation and test data to prevent data leakage. Error metrics were calculated in the outcome's original units to preserve interpretability. *MAE* was used to summarize typical error, while *RMSE* highlighted sensitivity to large deviations. R^2 and adjusted R^2 were reported for linear model performance and parsimony, using held-out predictions.

Residual distributions for the final model (XGBoost) were examined for centering and dispersion across both training and test data. Residuals-versus-predicted plots were used to detect heteroscedasticity. Because the dataset included repeated school-year observations, potential

optimistic bias in a random split remained a limitation. A grouped or time-aware data split was noted as a potential extension for future robustness checks.

Limitations

This study had several limitations due to the use of secondary data. First, several variables unrelated to dropout rates contained over 50% missing data and were excluded during data cleaning. While this improved analytic quality, it may have limited the range of potential predictors and reduced generalizability (Hegde et al., 2019). Second, some variables contained zero values whose interpretation was unclear. It could not be confirmed whether these zeros represented true values (e.g., zero dropouts) or missing/unreported data. This ambiguity may have affected both model accuracy and result interpretation. Third, the results may not generalize to all U.S. high schools due to contextual differences in geolocation, demographics, and socioeconomic conditions. Additionally, some variables were only reported at the district level, limiting school-specific analysis. To mitigate misinterpretation, detailed definitions of key terms were provided and applied consistently throughout the study.

Delimitations

This study was conducted using existing data from the ISBE. As such, the quality of the data was constrained by the validity and reliability of the instruments used by ISBE during data collection. The scope of the study was limited to high schools and school districts within the state of Illinois. Therefore, the findings may not be generalizable to all high schools and districts across the United States.

Only the following independent variables were used to predict high school dropout rates in Illinois: county, district type, school type, district size, grades served, number of student enrollment, rate of student enrollment White, rate of student enrollment Black or African

American, rate of student enrollment Hispanic or Latino, rate of student enrollment Asian, rate of student enrollment with disabilities, rate of student enrollment English learners, rate of student enrollment with low income, rate of student enrollment homeless, student attendance rate (*SAR*), student mobility rate (*SMR*), student mobility rate English learners (*SMREL*), student mobility rate with IEP, chronically truant students, student chronic truancy rate, high school dropout rate total, high school dropout rate male, high school dropout rate female, high school dropout rate White, high school dropout rate Black or African American, high school dropout rate Hispanic or Latino, high school dropout rate Asian, high school dropout rate English learners, high school dropout with IEP, high school dropout rate with low income, average class size high school, average number of days of physical education per week per student, teacher retention rate, and principal turnover within six years. Future researchers should investigate additional variables not addressed within the scope of this study. The timeframe analyzed in this study spanned from 2017 to 2019 and from 2020 to 2022.

Ethical Assurances

The existing data from the ISBE were not used in this study until the research methodology was approved by the Institutional Review Board (IRB) at National University. The reduced report card public dataset from ISBE, which was used to conduct this study, was stored securely in an offline location in a locked drawer, protected by firewalls and intrusion detection systems. Although this study will be published and both the data and code developed during the research will be made publicly available in a repository, publication guidelines will be followed to prevent any risk of data breach. This study did not involve human participants. The reduced report card public dataset from ISBE is anonymous; it does not contain individual names or identification codes. No informed consent or permission to use the data was necessary, as the

dataset is publicly available and unrestricted. There were no privacy concerns. Since the dataset is public and access-free, no confidential or personally identifiable information was used. Data collection and statistical analysis were conducted only after IRB approval was obtained. All IRB requirements were followed to ensure ethical and procedural compliance throughout the study.

Summary

This chapter covered the research methodology and design, population and sample, materials, operational definitions of variables, study procedures, data analysis plan, assumptions, limitations, delimitations, and ethical assurances. A quantitative, non-experimental research design was selected for this study to develop predictive models for high school dropout rates using demographic, geolocation, and internal school factors. The reduced Illinois report card public datasets for 2017, 2018, 2019, 2020, 2021, and 2022 were used.

The researcher analyzed data at both the school and district levels to examine the predictive relationships between the selected independent variables and the dropout rate. The models employed, multiple linear regression, random forest, and XGBoost regression were evaluated for their effectiveness in predicting dropout rates. In the next chapter, the findings and results of the study are presented.

Chapter 4 Results

This quantitative study employed multiple linear regression, random forest, and XGBoost models to identify the factors associated with high school dropout rates in Illinois, comparing the pre–COVID-19 period (2017–2019) with the COVID-19 period (2020–2022). By integrating demographic, geospatial, and school-level predictors, the models illuminated both linear and complex nonlinear relationships, providing actionable insights to reduce dropout rates. The researcher used data from 2017 to 2019 to represent the period before COVID-19 and from 2020 to 2022 to represent the COVID-19 period. Predictor variables included student demographics, geolocation, and other internal school factors. The annual high school dropout rate for each Illinois public high school (i.e., the percentage of students who left school without graduating in a given year) served as the dependent variable in all models. The Python codes used to perform the analysis for this study were saved in GitHub repository (https://github.com/claudengantchou/ClaudeNgantchou.git).

As little is known about the factors contributing to high dropout rates, this study explored the predictive relationship between demographic, geolocation, and internal school variables and high school dropout rates. Designing and validating this predictive model was essential for addressing the research questions and offering solutions to reduce the persistent problem of high school dropout.

Despite numerous studies examining dropout rates for general education students and students with disabilities, limited research has addressed the role of demographic, geolocation, and internal school factors in student dropout behavior. This study therefore provided insights that may be used to inform efforts to reduce dropout rates for both general education students

and students with disabilities in the state of Illinois. The study was guided by two research questions (RQs):

RQ1: To what extent, if any, do the demographic, geolocation, and other internal factors of the schools predict the high school dropout rate in the state of Illinois?

RQ2: Which predictor variables significantly influence the prediction of the high school dropout rate in the state of Illinois?

This chapter embodied the modeling phase of the CRISP-DM framework by outlining, in sequence, the practical steps taken to build and evaluate the predictive models. It begins with data preparation and cleaning, during which the researcher imputed missing values and addressed anomalies. This was followed by the construction of a correlation matrix and related feature-engineering decisions to reduce multicollinearity and enhance interpretability. A brief descriptive analysis explored key distributions and relationships prior to modeling.

The predictive modeling section documents how multiple linear regression, random forest regression, and XGBoost regression were trained using an 80/20 train–test split. A dedicated hyperparameter tuning and model evaluation segment details the grid search procedures, cross-validation results, and comparative performance metrics (R^2, *MAE*, *RMSE*). The feature importance analysis interprets the XGBoost output and identifies the most influential predictors of dropout rates. A residual analysis for the final model assesses error distributions and confirms that residuals were broadly centered around zero with no systematic patterns.

Data Collection

The researcher collected data from public high schools in Illinois. The source of the data was the ISBE. The period of data collection spanned from 2017 to 2022 and included aggregated data on high school dropout rates and school-level demographics, geolocation, and internal

factors. Although the ISBE dataset also contained information on charter, elementary, middle, and pre-K schools, only high school data were used in this study.

Data Preparation and Cleaning

The datasets received from ISBE were organized by reporting year. A total of 23,110 cases were retrieved from the six datasets. The distribution was approximately equal across years (about 16% each). The researcher removed variables unrelated to the study's focus and excluded records for non–high school institutions to isolate the high school population. High schools accounted for 16.8% (n = 3,892) of the total dataset and formed the final analytic sample.

Of the 35 variables initially selected, the researcher removed 18 due to high levels of missing data. Most of the excluded variables exceeded the acceptable missingness threshold of 50%. These included demographic variables (e.g., enrollment by race/ethnicity), as well as variables related to English learners, homeless students, students with IEPs, and subgroup-specific dropout rates (e.g., for Asian, Black, Hispanic, White, male, female, and low-income students). For example, the mobility rate for English learners had a missingness rate of 67.0%, and the dropout rate for Asian students was missing in 61.4% of records. These variables were excluded to preserve model integrity and reduce potential bias or instability in the predictions.

The researcher conducted descriptive analysis on the remaining variables to assess levels of missing values, outliers, and skewness in the data. The Shapiro–Wilk test of normality on the individual variables indicated that all variables were not from a normally distributed population ($p < .05$). Several variables also exhibited skewness values greater than 1, exceeding the threshold typically associated with normal distributions. Based on these findings, the researcher imputed missing values using the median for numeric variables. For categorical (object) variables, the mode (most prevalent value) was used. Outliers were detected using the z-score

method (standardization), which flagged values more than 3 standard deviations from the mean. Nearly all variables contained outliers, with the exception of three variables.

Label Encoding

The next phase involved data transformation through label encoding, in which categorical variables were converted into numerical format for machine learning analysis. Although five categorical variables were available, only one was successfully transformed.

- *School type* was excluded, as the dataset already included only high schools and the variable served as an eligibility filter rather than an independent predictor.
- *District type* shared similar values with *school type* (e.g., "high school") and was excluded to avoid redundancy.
- *Grades served* was a multiple-response variable and could not be transformed without losing structural integrity, as it would require representing multiple values per case.
- *County* was a nominal variable with over 100 categories. Label encoding this variable would imply an ordinal relationship (e.g., $0 < 1 < 2$), which could mislead machine learning models.

The only categorical variable that was successfully label encoded was *district size*, which classified districts as small, medium, or large.

COVID-19 Period Variable and Scaling

To investigate differences in variable behavior before and during COVID-19, the researcher created a categorical variable to define the COVID-19 period: 2017 to 2019 as pre-COVID and 2020 to 2022 as during COVID. This variable was generated based on the reporting year in the datasets.

Subsequently, the researcher scaled the data using Python's StandardScaler from sklearn.preprocessing, standardizing numerical features to a mean of 0 and a standard deviation of 1. Histograms revealed continued issues with distribution, and boxplots showed that several outliers remained.

Transformation

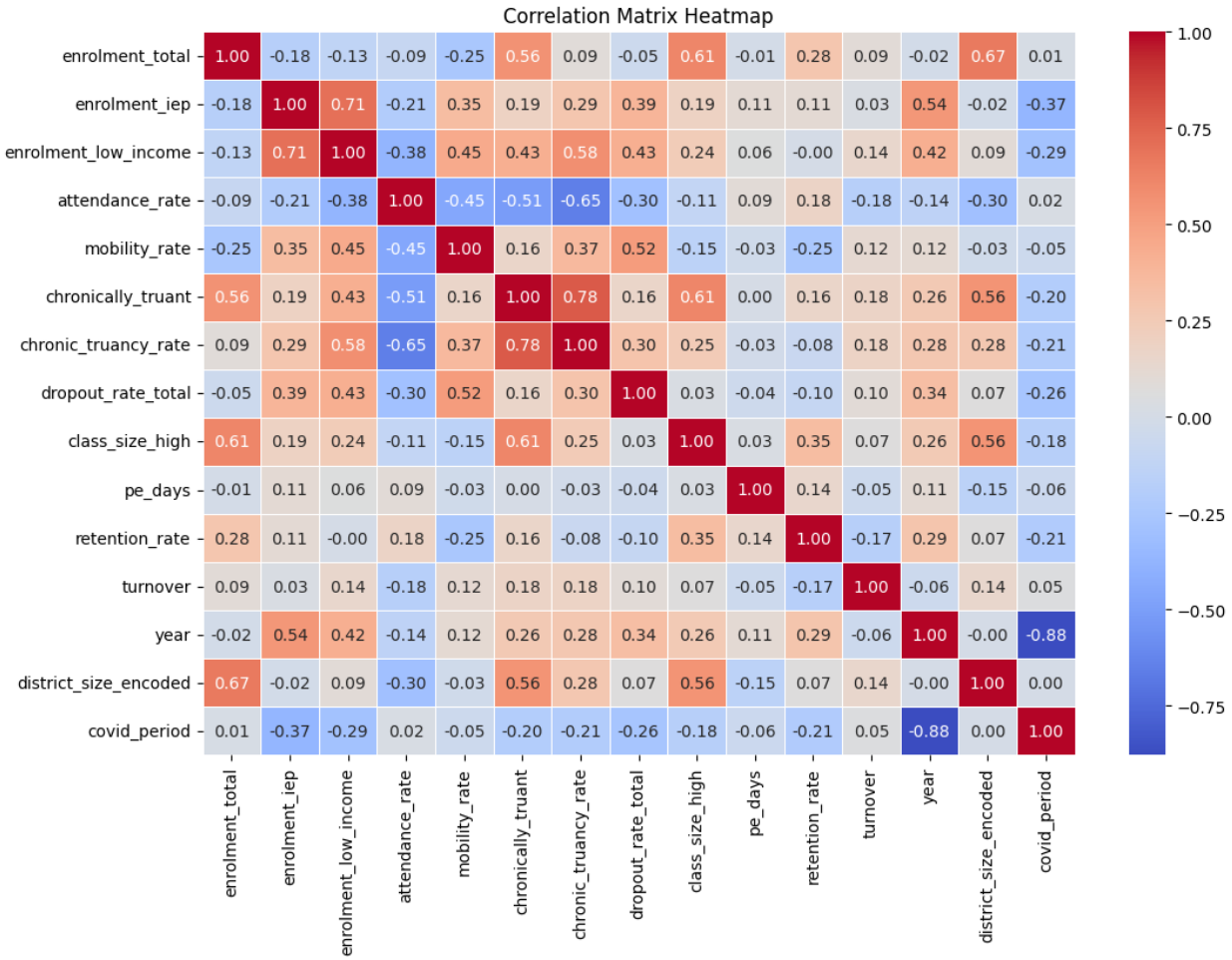
Given the presence of skewness and outliers (as indicated by histograms and descriptive statistics), the researcher applied transformations to improve normality. While most variables exhibited skewness within the acceptable ±1 range, several exceeded this threshold. Because the data included both positive and negative values, the Yeo–Johnson transformation was implemented (Hamasha et al., 2022). Unlike the Box–Cox method, Yeo–Johnson can normalize variables regardless of sign and is effective in stabilizing variance. This transformation improved distribution characteristics, as evidenced by reductions in skewness and more symmetric histograms.

Correlation Matrix

The final step in data preparation involved examining the correlation matrix to identify variable pairs with strong correlations that could indicate multicollinearity. As shown in the heatmap (see Figure 3), the variable *year* exhibited a strong negative correlation with *covid_period* ($r = –.88$), indicating that both variables captured similar temporal dynamics. To avoid redundancy, *year* was excluded, and *covid_period* was retained for its clearer interpretability in distinguishing pre- and during-COVID years. No other variable pairs showed correlations at or above the exclusion threshold ($|r| \geq .70$). Therefore, all remaining variables were retained for further analysis.

Figure 3

Correlation Matrix Heat Map



Feature Engineering

The researcher conducted feature engineering by addressing highly correlated features that could introduce multicollinearity and potentially create redundant information in the model. The variance inflation factor (VIF) was used to detect multicollinearity. VIF values indicate how much a predictor variable is correlated with other predictors. Values above 10 are commonly considered to reflect significant multicollinearity (Kalnins & Praitis-Hill, 2023). Variables with such values should typically be dropped or combined.

The variable *chronically truant* had a VIF of 9.595, which, although slightly below the threshold, indicated a high degree of multicollinearity. The researcher removed this variable to reduce redundancy and improve model interpretability. High VIF values suggest that a variable is strongly correlated with other predictors, which can distort regression coefficients and reduce reliability.

The variable *year* also showed a VIF of 6.761, indicating a moderate level of collinearity. Although this was below the critical threshold, the researcher chose to remove it based on both statistical evidence and theoretical justification. The *covid_period* variable, which provided a clearer categorical distinction between the pre- and during-COVID-19 periods, was retained in its place.

Removing these two variables helped mitigate multicollinearity and improved the stability of the regression model. After removal, the VIF values for all remaining variables fell below the threshold of 10, ensuring that the model's estimates were not influenced by redundancy. Although removing *chronically truant* and *year* may have slightly limited the model's sensitivity to truancy and temporal nuances, these decisions enhanced the reliability and interpretability of the results.

Regularization and Feature Importance

Following the removal of multicollinear variables, the researcher used Elastic Net regression to further evaluate residual collinearity and identify feature importance. Elastic Net combines L1 (Lasso) and L2 (Ridge) regularization to penalize and potentially eliminate weak predictors, thereby stabilizing the model and highlighting influential variables (Mohanasundaram & Rangaswamy, 2025).

Elastic Net identified the following as strong predictors: *mobility rate*, *enrolment low income*, *enrolment total*, and *covid_period*. Several other variables received zero or near-zero coefficients. However, because all remaining variables had acceptable VIF values and were theoretically relevant to the research, none were removed. Instead, the Elastic Net analysis was used to rank predictor importance, allowing the researcher to preserve the interpretive completeness of the model while still reducing the risk of multicollinearity.

Results

Descriptive Analysis

Table 5 includes descriptive statistics for the primary study variables. The dataset included educational and demographic characteristics from 3,892 high schools in Illinois. A plurality of the high schools were classified as large (n = 1,692, 43.5%), followed by medium-sized schools (n = 1,497, 38.5%), see Figure 4. Small-sized schools were the least represented, comprising 17.0% of the total (n = 661). Enrollment numbers varied widely across schools, ranging from 12 to 4,613 students (M = 878.2, SD = 883.0). High school dropout rates were low, with a mean of 3.2% (SD = 3.9). As for teacher features collected in the study, teacher retention rates were high, with a mean of 85.7% (SD = 10.5), and principal turnover within six years averaged 2.0 (SD = 1.2).

Table 5

Descriptive Analysis

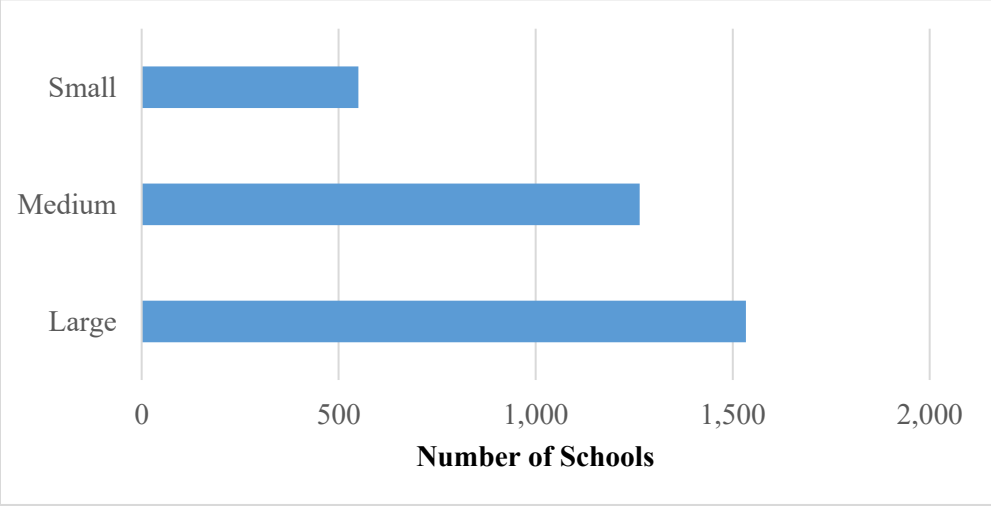
	N	*Min*	*Max*	*M*	*SD*	*Mdn*
Enrolment	3,892	12.00	4613.00	898.17	882.98	473.00

Individualized Education Programs (%)	3,892	0.00	100.00	12.15	7.91	12.60
Low Income (%)	3,892	0.00	100.00	38.47	27.60	36.60
Attendance Rate	3,892	51.90	100.00	91.86	5.22	93.30
Student Mobility Rate	3,892	0.00	100.00	10.0	9.53	8.10
Student Chronic Truancy Rate	3,892	0.00	100.00	17.32	21.26	9.40
High School Dropout Rate	3,892	0.00	52.00	3.24	3.89	2.30
Class Size	3,892	2.50	33.50	17.86	5.04	17.40
Number of Days of Physical Education per Week Per Student	3,892	0.00	5.00	4.45	1.14	5.00
Teacher Retention Rate	3,892	34.30	100.00	85.65	10.68	88.40
Principal Turnover within 6 Years	3,892	1.00	21.00	2.01	1.19	2.00

Note. *n* = Sample size; *Min* = Minimum; *Max* = Maximum; *M* = Mean; *SD* = Standard deviation; *Mdn* = Median

Figure 4

School Types



Predictive Modeling

The researcher used multiple linear regression, random forest, and XGBoost regression models to analyze the impact of various factors on high school dropout rates. Prior to model fitting, all variables were reviewed and recoded to ensure that the analytic matrix contained only numeric values. The dependent variable, the annual dropout rate, remained a continuous proportion (0–100%). Nominal categorical predictors such as *district size* (small, medium, large) and *covid_period* (pre-COVID-19 and during COVID-19) were transformed using one-hot

encoding. This procedure converted each category into a binary numeric indicator, allowing inclusion in all three models without violating assumptions. The total dropout rate served as the target variable, and all other eligible variables were selected as independent predictors, excluding those removed due to multicollinearity or excessive missingness.

The dataset was split into training and testing subsets using an 80/20 ratio. Eighty percent of the data were used to train the models, and 20% were used for testing. A random seed was set to ensure reproducibility. The training data were standardized through fitting and transforming, while the testing data were only transformed to avoid data leakage (Rosenblatt et al., 2024). The researcher trained a multiple linear regression model on the scaled training data and used it to predict dropout rates in the test data. Random forest and XGBoost regression models were also trained using the same training data, each configured with 100 estimators and a fixed random seed.

Model performance was evaluated using three metrics: the coefficient of determination (R^2), mean absolute error (*MAE*), and root mean squared error (*RMSE*). The R^2 score measured the proportion of variance in the dependent variable explained by the model (Chicco et al., 2021). *MAE* represented the average absolute difference between predicted and actual values (Balawi & Tenekeci, 2024). *RMSE* was the square root of the average squared prediction errors, penalizing larger errors more heavily (Palma et al., 2025). Mean absolute percentage error (*MAPE*) was not used due to its instability in the presence of zero or near-zero actual values. Because *MAPE* involves division by the actual value, its use can lead to undefined or distorted results (Iida, 2024).

Model Performance Results

Table 6 presents the performance results for the multiple linear regression, random forest, and XGBoost regression models. Among the three, XGBoost yielded the best overall performance. It achieved the highest R^2 score of .467, indicating that it explained approximately 46.7% of the variance in high school dropout rates. The random forest model followed, explaining 45.8%, and the multiple linear regression model explained 30.2%. A higher R^2 reflects a better model fit, indicating that the ensemble models better captured the structure of the data.

For *MAE*, XGBoost again outperformed the other models with the lowest error (0.546), followed by random forest (0.555) and multiple linear regression (0.649). *RMSE* was also lowest for XGBoost (0.729), followed by random forest (0.735), and multiple linear regression (0.834). Because *RMSE* penalizes larger errors more heavily, these results reinforce the conclusion that the ensemble models generated more accurate predictions.

Overall, both random forest and XGBoost outperformed multiple linear regression across all evaluation metrics. XGBoost consistently delivered the best performance, indicating it was the most suitable model for predicting high school dropout rates in this study. These findings support the use of advanced ensemble learning methods in educational data mining, particularly when data includes nonlinear relationships and interaction effects among predictors.

Table 6

Model Performance Results

Model	R^2 Score	Mean Absolute Error (*MAE*)	Root Mean Squared Error (*RMSE*)

Multiple Linear Regression	0.302	0.649	0.834
Random Forest	0.458	0.555	0.735
XGBoost	0.467	0.546	0.729

Feature Importance Analysis

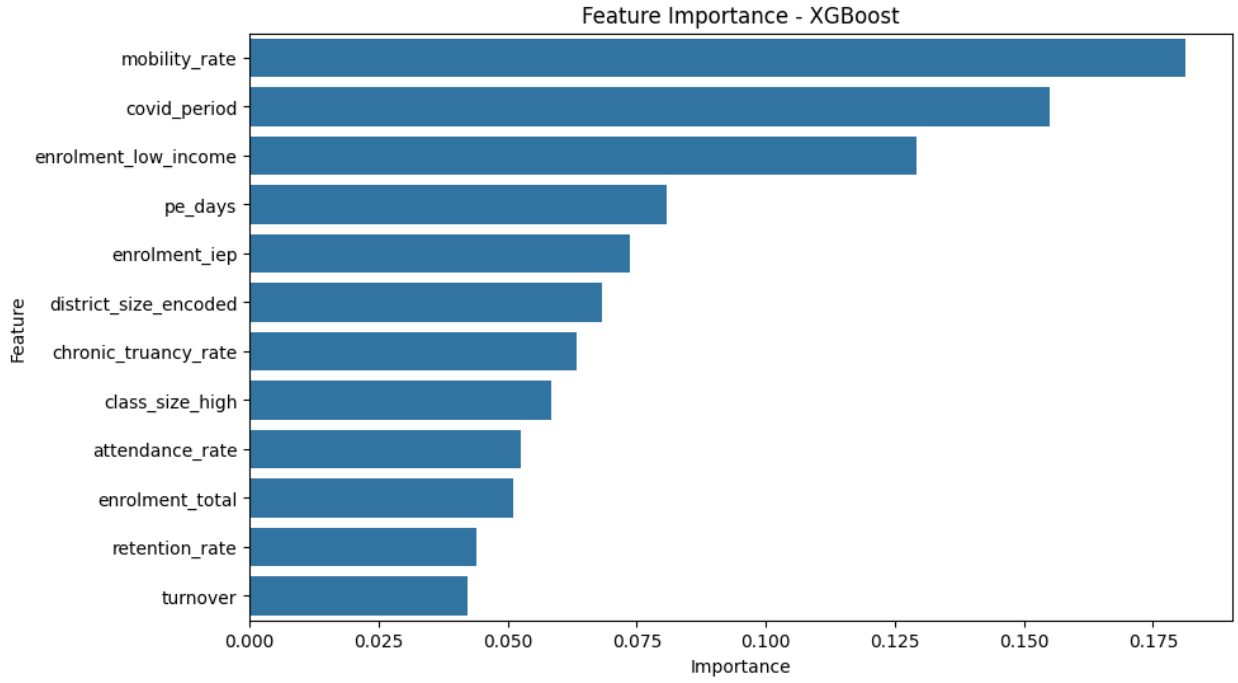
Because the XGBoost regression model demonstrated the highest predictive performance among the three models, the researcher conducted an exploratory analysis to examine the relative importance of the input features in predicting high school dropout rates. Figure 5 presents the feature importance scores derived from the XGBoost model.

The analysis revealed that *student mobility rate* was the most influential predictor, indicating that schools with high levels of student turnover or instability tended to experience elevated dropout rates. The second most important feature was *covid_period*, suggesting a strong temporal effect on dropout patterns, likely tied to learning disruptions during the COVID-19 pandemic. Enrollment of *low-income students* also emerged as a key predictor, underscoring the role of socioeconomic disadvantage in student persistence.

Other notable predictors included *PE days*, *enrollment in Individualized Education Programs (IEPs)*, and *district size*, which may reflect variations in school resources, support services, and administrative structure. Additionally, *chronic truancy rate*, *class size*, and *attendance rate* were ranked moderately high in importance, reinforcing the well-documented connection between school engagement and dropout risk. Less influential predictors, such as *turnover rate*, *total enrollment*, and *retention rate*, still contributed to the model but demonstrated comparatively lower predictive power.

Figure 5

XGBoost Feature Importance



Hyperparameter Tuning and Final Model Evaluation

The researcher performed hyperparameter tuning on all three models, multiple linear regression, random forest, and XGBoost, to improve predictive accuracy and determine optimal configurations for deployment. Each model was evaluated using the standard performance metrics: R^2, mean absolute error (*MAE*), and root mean squared error (*RMSE*).

The random forest model was tuned using a predefined grid of hyperparameters. Key parameters included the number of estimators (trees), maximum tree depth, minimum samples required to split a node, minimum samples required to be a leaf, and the number of features considered at each split. After fitting various combinations to the training data, the optimal configuration was identified: 500 estimators, maximum depth = 20, *min_samples_split* = 5, *min_samples_leaf* = 4, and the *sqrt* strategy for max features. The tuned model achieved reduced predictive accuracy with an R^2 of .452, *MAE* of .569, and *RMSE* of .739. These results showed

no meaningful gains over the untuned model and indicated that hyperparameter optimization did not improve model performance.

The XGBoost model was also tuned using a grid search approach. Key hyperparameters included learning rate, maximum tree depth, number of estimators, and subsample ratio. The best configuration consisted of 100 estimators, learning rate = .05, maximum depth = 7, and subsample = .8. The tuned model achieved an R^2 of .466, *MAE* of .555, and *RMSE* of .729. Although the performance improvement was modest, the XGBoost model remained competitive and performed nearly as well as the tuned random forest model. It also showed slightly better *MAE* and stronger generalization due to its regularized boosting framework.

To strengthen the baseline multiple linear regression, the researcher implemented Ridge regression with standardized inputs using `StandardScaler`. The regularization parameter (α) was tuned to identify the best-performing model. An α value of 10 produced an R^2 of .302, *MAE* of .649, and *RMSE* of .834. While Ridge regression did not match the accuracy of the ensemble models, it offered a simple, interpretable, and computationally efficient benchmark.

Overall, hyperparameter tuning did not yield consistent performance improvements across models. The tuned random forest model performed substantially worse than its untuned counterpart, indicating that optimization reduced predictive accuracy rather than enhancing it. For XGBoost, tuning produced performance metrics that were essentially identical to those of the untuned model, suggesting limited sensitivity to further optimization within the explored parameter space. Despite these findings, XGBoost remained the best-performing model overall, achieving the highest R^2 and the lowest error metrics among all models evaluated. While the ensemble methods outperformed linear and regularized regression in general, the results suggest that model choice was more influential than hyperparameter tuning in this analysis.

Consequently, XGBoost was selected as the final predictive model due to its strong baseline performance, robustness, and capacity to capture complex relationships associated with high school dropout rates.

Final Feature Importance: XGBoost

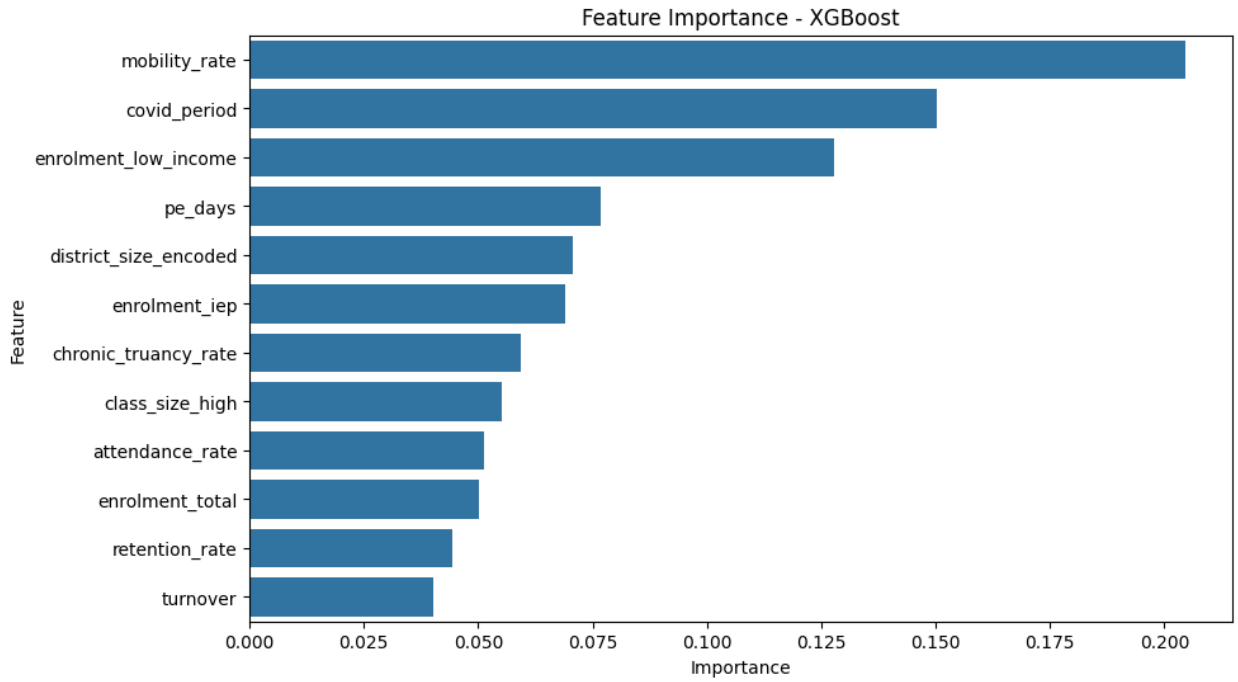
The researcher used the final tuned XGBoost regression model to identify the most influential predictors of high school dropout rates in Illinois. Figure 6 presents the resulting feature importance scores. These scores were derived from XGBoost's built-in feature-importance metric (not SHAP values) and should be interpreted as relative contributions within the fitted model. The top three predictors were *mobility rate*, *covid_period*, and *low-income student enrollment*. These results highlighted the critical role of both structural school-level factors and broader contextual conditions in shaping dropout risk. *Mobility rate* emerged as the strongest predictor, reinforcing prior research that identified student transience and school instability as major contributors to academic disengagement and attrition.

Covid_period ranked second in importance, suggesting that the pandemic had a substantial effect on dropout patterns, likely tied to disrupted learning environments, limited student support, and elevated socioeconomic stress. *Low-income student enrollment*, *IEP enrollment*, and *PE days* also contributed significantly, reflecting the combined impact of economic hardship, special education needs, and access to whole-child programming.

Variables, such as *district size*, *attendance rate*, and *retention rate* were comparatively less influential but still contributed to the model's predictive performance.

Figure 6

XGBoost Feature Importance (Final Tuned Model)



Residual Analysis for the XGBoost Model

To assess model fit and error behavior for the final XGBoost regression model, the researcher performed a residual analysis by examining the distribution of residuals (i.e., the differences between predicted and actual dropout rates) for both the training and testing datasets. Figure 7 displays histograms of the residuals for each set, with superimposed density curves and vertical reference lines at zero to assist visual interpretation.

The residuals from the training set (left panel) exhibited a roughly normal distribution, symmetrically centered around zero. This indicated that the model's predictions on the training data were generally unbiased and that the errors were evenly distributed, with most clustering near zero. The bell-shaped distribution further suggested that the model did not systematically over- or under-predict on the training data.

In contrast, the residuals from the test set (right panel) were less symmetric and more dispersed, showing mild skewness and a flatter, multimodal shape. Although the test residuals remained broadly centered around zero, the wider spread indicated increased variability in prediction errors on unseen data. This pattern suggested that, although the model did not exhibit severe overfitting, its accuracy and consistency were reduced when generalizing beyond the training data.

Figure 7

Training and Test Residuals Distribution

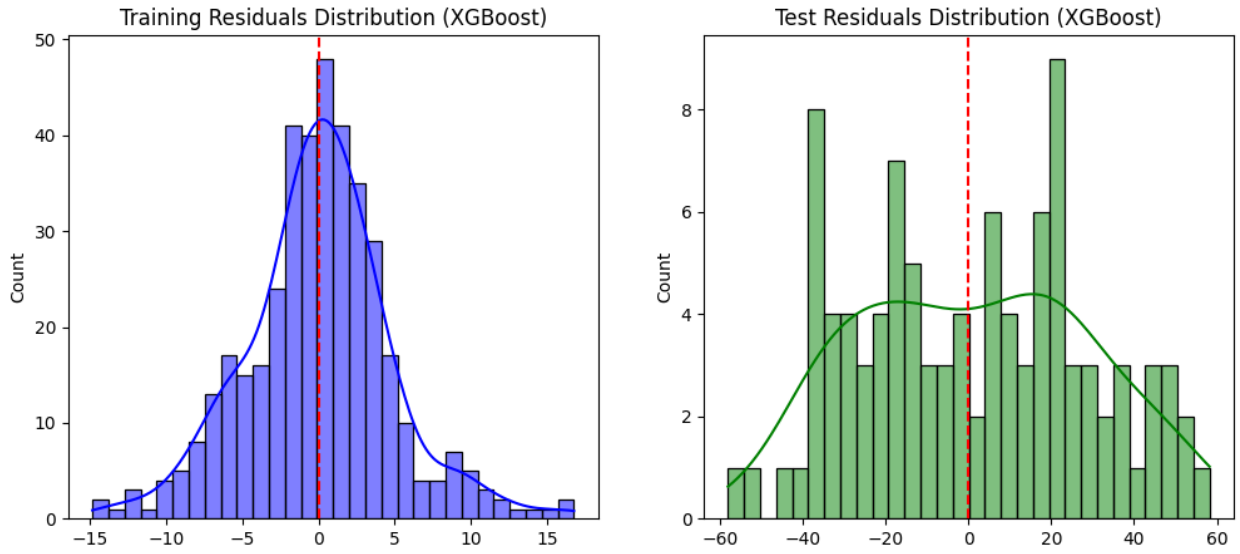
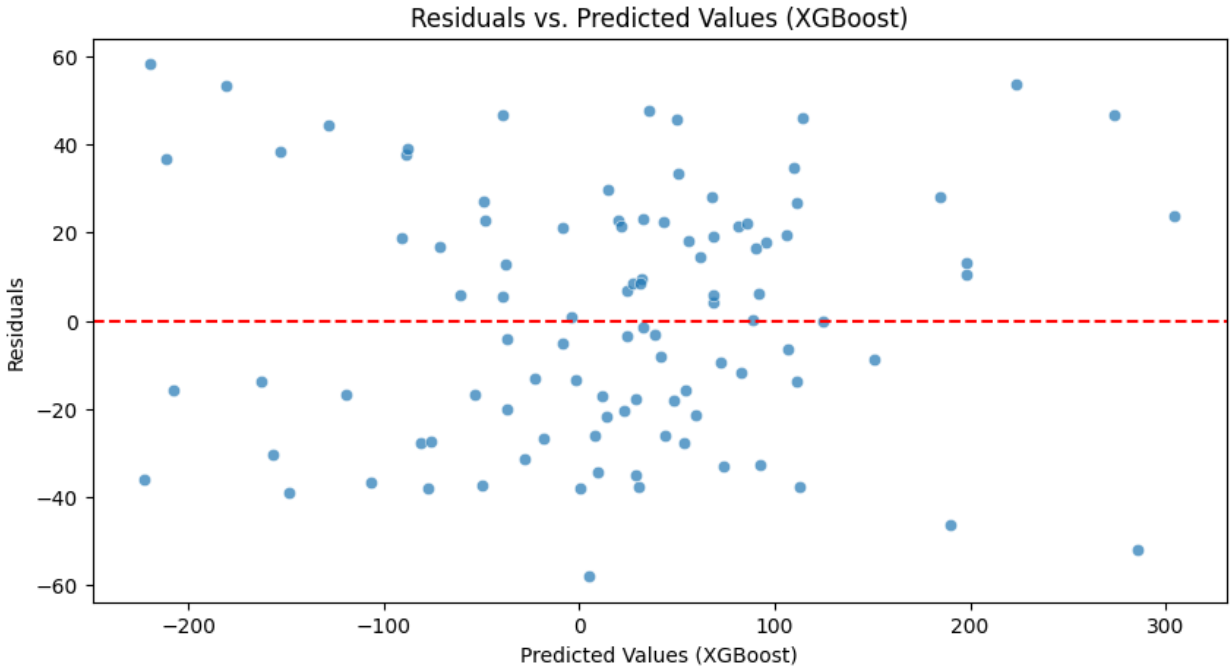


Figure 8 presents the scatter plot of residuals versus predicted values for the final XGBoost regression model. This diagnostic plot was used to evaluate whether the model's errors were randomly distributed and to detect potential violations of model assumptions, such as non-linearity or heteroscedasticity. The residuals appeared randomly scattered around the horizontal zero line, with no clear or systematic pattern. This distribution suggested that the model's errors were approximately homoscedastic, meaning the variance of the residuals remained relatively constant across the full range of predicted dropout rates. The absence of funnel-shaped or curved

patterns implied that the model was reasonably well-fitted and did not systematically under- or over-predict at specific predicted value intervals.

Figure 8

Residuals vs Predicted Values Scatter Graph



Integrating Feature Importance with the Theoretical Framework

This chapter focused on developing and evaluating predictive models to examine factors influencing high school dropout rates in public schools across Illinois. Three supervised machine learning models, multiple linear regression, random forest regression, and XGBoost regression, were designed, tuned, and evaluated. The models were trained on school-level data incorporating demographic, academic, and institutional features. Model performance was evaluated using R^2, mean absolute error (*MAE*), and root mean squared error (*RMSE*). Among the models, XGBoost demonstrated the most accurate and consistent predictive performance, achieving the lowest

MAE and a strong R^2 score. In addition, XGBoost offered a superior balance of accuracy and efficiency, making it the preferred model.

The feature importance plot from the final XGBoost model (Figure 6), based on the model's built-in importance metric, showed that *mobility rate* was the most influential predictor, followed by *COVID-19 period*, *low-income student enrollment*, and *IEP enrollment*. These findings underscored the importance of both structural school-level factors and contextual disruptions. Specifically, the results highlighted how student transience, economic disadvantage, special education needs, and the COVID-19 pandemic shaped dropout risk.

Although earlier model versions considered additional subgroup-specific dropout rates (e.g., by gender), these features were excluded due to high levels of missing data. Contrary to initial assumptions, the *COVID-19 period* variable emerged as one of the top predictors, indicating its significant impact on dropout patterns and challenging earlier beliefs that its effect was marginal.

The overall findings suggest that addressing systemic inequities, such as poverty and special education access, alongside contextual challenges like pandemic-related disruptions is essential for reducing dropout rates across Illinois public schools.

Chapter 5: Implications, Recommendations, and Conclusions

The problem addressed in this study was the high rate of high school dropout in the state of Illinois from 2017 to 2022. The purpose of this quantitative study was to determine a predictive model for high school dropout rates before and during COVID-19 in Illinois, from 2017 to 2019 and from 2020 to 2022, respectively. Predictor variables included student demographics, geolocation, and other internal school-level factors. The response variable was the annual high school dropout rate. Because little was known about the predictive factors associated with elevated dropout rates, the researcher explored the relationship between the selected predictors and dropout outcomes. Designing and validating this predictive model addressed the study's research questions and contributed actionable insights toward reducing dropout risk.

Building from these anchors, this chapter synthesizes the study's methodology, design, results, and limitations before turning to implications, recommendations, and conclusions. Because the investigation was explicitly framed around machine-learning prediction of dropout rates in pre- and during-COVID-19 conditions in Illinois, all interpretations are tied directly to that scope and the guiding research questions.

Methodologically, the researcher employed a quantitative, non-experimental predictive modeling design using publicly available secondary data from the Illinois Report Card at the school and district levels. Multiple linear regression (MLR), random forest (RF), and XGBoost regression models were specified, trained, tuned, and evaluated using standard out-of-sample procedures (i.e., 80/20 splits and cross-validation), and assessed using performance metrics suitable for continuous outcomes (R^2, *MAE*, *RMSE*). This approach allowed for comparison of both linear and non-linear relationships between predictor variables and annual dropout rates.

In Chapter 4, the XGBoost model presented itself as the best overall model ($R^2 = 0.467$) and also yielded the lowest MAE (0.546) and RMSE (0.729). Random forest took second place, followed by multilinear regression (MLR). These results give strong evidence that tree-based ensemble models were a better paradigm for prediction on this dataset. In addition, feature importance from the final XGBoost model indicated that the student mobility rate was the most important predictor, followed by the COVID-19 period, and low-income student enrollments. IEP enrollments and days of physical education (PE) were also moderate but important predictors. These results highlighted the importance of school stability, pandemic-related disruptions, and levels of socioeconomic disadvantage in predicting variation in dropout across Illinois high schools within the study period.

In line with the predictive model discussed in Chapter 3, the analysis pipeline was meant to balance interpretability and predictive accuracy by comparing parametric and ensemble methods while still adhering to described data preparation and validation processes. A brief review of limitations is necessary to frame the upcoming discussion. As mentioned above, the use of secondary data necessitated exclusion of variables with a considerable degree of missingness (e.g., >50%) in certain instances, which narrowed the possible scope of predictors and consequently may have diminished the generalizability we sought in the study. The researcher applied out-of-sample validations but did not account for the possibility that being premised on standard random splits there may have been a grouping or time-aware data split to allow for potentially even more robust information on school or time level generalizability. Finally, interpretability varied across models; for example, linear regression is more transparent than boosted trees.

These data and design realities must be considered when translating findings into implications for policy and practice. The remainder of this chapter discusses the implications in light of the research questions and literature reviewed in Chapter 2. Recommendations for practice and future research follow, concluding with a summary of the study's key contributions.

Implications

Theoretical Implications

This study indicates that school-level conditions, especially student mobility, the COVID-19 period, and low-income enrollment, are systematically associated with variation in annual high school dropout rates. Interpreted via the study's survival-analysis framing, both persistent exposures (e.g., concentrated poverty, mobility) and acute shocks (e.g., pandemic disruption) plausibly shift the baseline hazard of dropout at the school-year level (e.g., Bauman & Cranney, 2020; Grigg, 2012; Rumberger, 2020; Seymour et al., 2020). With XGBoost $R^2 \approx 0.467$ on held-out data, the models exhibit moderate explanatory power suitable for risk screening and planning, while acknowledging substantial unexplained variance likely tied to individual, family, and classroom dynamics (consistent with the multilevel accounts summarized in Chapter 2). These findings extend the literature by grounding abstract constructs, stability, socioeconomic disadvantages, and system disruption, in an empirically ranked set of predictors from modern ensemble learners.

Methodological Implications

Methodologically, comparing multiple linear regression, random forest, and XGBoost shows that tree-based ensembles better capture nonlinearities and interactions present in school-level data than linear baselines, whereas linear regression remains useful for transparent coefficient interpretation and benchmarking (Martínez-Plumed et al., 2019; Schröer et al., 2021).

Out-of-sample evaluation (80/20 split, cross-validation) and residual diagnostics justify treating model outputs as probabilistic decision support, not deterministic classifications. Preprocessing choices that improved stability, excluding variables with >50% missingness and pruning collinear fields, also likely attenuated some relationships, underscoring the value of time-aware validation, replication in future cohorts, and improved data completeness (especially subgroup indicators) in subsequent studies. Finally, feature-importance summaries from ensembles should be interpreted as descriptive (non-causal) signals of relative predictive contribution, which aligns with the study's non-experimental design (Xing & Du, 2019).

Practical and Policy Implications

Given the models' scope and performance, the most appropriate applications are screening, triage, and planning at the district/school level, paired with educator judgment**.**

1. **Stabilize Enrollment and Transitions (Mobility as top signal).**
    - **District policy:** Adopt cross-district records-continuity and credit-transfer protocols; require rapid onboarding (placement, schedule, service continuity) within 5 school days for new transfers.
    - **School practice:** Stand up a mobility team (counselor, registrar, case manager) with a 30-day academic/attendance check-in for all incoming students (Grigg, 2012; Rumberger, 2020).
2. **Cushion Economic Strain in High-Poverty Zones (Low-income enrollment).**
    - **District policy:** Target weighted funding and embedded supports (on-site tutoring, mentoring, family navigation) to schools above poverty thresholds; conduct quarterly equity audits of support access.

- **School practice:** Integrate basic-needs referrals into Multi-Tiered System of Supports (MTSS) teams; pair flagged cohorts with attendance supports and credit-recovery options (Rumberger, 2020).

3. **Continuity and Re-engagement (COVID-period as system shock).**
   - **District policy:** Maintain an annually exercised Continuity of Learning Plan (remote/hybrid pivots, device/connectivity, re-entry credit audits).
   - **School practice:** After closures/absence clusters, run targeted re-engagement (counselor outreach, schedule adjustments, credit audits) prioritizing cohorts surfaced by the model (Bauman & Cranney, 2020; Seymour et al., 2020).
4. **Disability-Responsive Retention (IEP enrollment).**
   - **District policy:** Require mid-year IEP transition checks for mobile students; guarantee flexible credit pathways and coordinated academic-behavioral supports.
   - **School practice:** Trigger IEP review and case management when risk indicators compound (mobility, truancy, and course failures) (Irwin et al., 2021).
5. **Data-Informed Early Warning with Guardrails ($R^2 \approx 0.47$).**
   - **District policy:** Deploy dashboards that combine model flags with simple, interpretable baselines (attendance, course failures), and require human-in-the-loop decisions to avoid punitive automation.
   - **School practice:** Use weekly lists to offer supports (tutoring, counseling, schedule fixes); monitor false positives/negatives and localize thresholds (Kearney & Childs, 2023; Rumberger, 2020; Xing & Du, 2019).
6. **Use Program-Adjacent Signals Diagnostically (PE days).**

   Treat PE days and similar features as proxies for engagement/climate, prompts for

auditing extracurricular access, advisory quality, and adult-student connectedness, rather than as direct levers (Rumberger, 2020).

Scope and Proportionality. Because models operate on school-year aggregates and explain a moderate share of variance, actions above should be framed as risk-aligned supports, then evaluated for equity and impact over time; this stance is consistent with the predictive (non-causal) design and the literature on mobility, poverty, and pandemic-era disruption (Bauman & Cranney, 2020; Grigg, 2012; Rumberger, 2020; Seymour et al., 2020).

Positive Social Change and Practice Implications (Probable Vs. Improbable). Because mobility and low-income enrollment emerge as top drivers, the most probable practice implications involve stabilizing enrollment and addressing resource strain. Credible strategies include coordinated transfer protocols, rapid onboarding supports for newcomers, records continuity across districts, and expanded school-embedded services in high-poverty zones (Rumberger, 2020). The strong *COVID-19* period signal supports maintaining contingency plans for instructional continuity, targeted counseling, and re-engagement after disruptions, so that future shocks do not lead to permanent exits. This is consistent with pandemic-era research summarized in Chapter 2 (Bauman & Cranney, 2020; Seymour et al., 2020).

Given the demonstrated predictive signal and prior evidence for analytics-enabled triage, districts should strengthen data-driven early warning and attendance-monitoring systems that proactively route supports (Kearney & Childs, 2023; Rumberger, 2020; Xing & Du, 2019). By contrast, it would be improbable, and beyond the present evidence, to claim that adjusting a single program input such as *PE days* could materially reduce dropout risk. The model suggests that such inputs are secondary and likely function as correlates of broader engagement or climate conditions.

Interpretation Boundaries

Several factors influence the interpretation of results. The models were trained on school-level aggregates from 2017–2022, allowing statewide school-year inferences but not insights into individual-level motivations or within-school variation, limitations already addressed in the methods chapter. Variables with high missingness were excluded, and overlapping fields were pruned to ensure model stability; while these decisions enhance generalization, they may also attenuate the observed influence of constructs imperfectly captured in the source data.

How to Read the Statistical Outputs Used Here

In this study, R^2 summarizes how much of the between-school variation in annual dropout rates the model accounts for on held-out data. The final XGBoost model achieved $R^2 \approx 0.467$ with the lowest MAE and RMSE among candidates, which we interpret as moderate explanatory power appropriate for screening and planning (not deterministic prediction). In the multiple linear regression baseline, coefficients quantify the expected change (in percentage points of the dropout rate) associated with a one-unit increase in each predictor holding other variables constant; these signs and magnitudes support narrative statements about directionality, while acknowledging the linear model's lower overall fit relative to the ensembles. For the tree-based models (Random Forest, XGBoost), we rely on feature importance scores (gain- or impurity-based) to indicate each predictor's relative contribution to out-of-sample prediction; these are descriptive, not causal. Consistent with Chapter 4, student mobility rate ranked as the top signal, followed by the COVID-19 period indicator and low-income enrollment, with additional contributions from IEP enrollment and PE days—patterns that align with the chapter's conclusions about structural instability, pandemic-era disruption, and socioeconomic disadvantage elevating dropout risk. Together, these statistics explain why we characterize the

models' performance as practically useful for triage and why we emphasize probabilistic, not causal, interpretation.

Although the researcher used held-out evaluation to guard against overfitting, residual diagnostics revealed broader test-set error dispersion than was present in training. This finding supports the use of predictions as probabilistic estimates rather than deterministic classifications. Replication using future cohorts or time-aware validation designs is advisable. Within these limits, the central implication remains consistent with Chapter 2: structural instability, pandemic-era disruption, and socioeconomic disadvantage jointly elevate dropout risk, and predictive models can surface these signals at a scale useful for triage and planning (Bauman & Cranney, 2020; Grigg, 2012; Rumberger, 2020; Seymour et al., 2020).

Recommendations for Practice

This section contains a few recommendations for practice that stem from the findings in Chapter 4. The recommendations are centered around stabilizing enrollment, supporting high-poverty schools, planning for disruptions in regular services, enhancing disability-responsive services, operationalizing early warning systems and responsibly interpreting proxy characteristics.

Stabilize Enrollment and Smooth Transitions

Since the student mobility rate was the most significant predictor of dropout risk, districts should treat stabilizing enrollment as a core prevention strategy. Specific action steps can include: (1) formalizing district transfer protocols; (2) expediting onboarding to allow new students rapid access to classes; (3) operationalizing a cross-district system for records to facilitate course placement, credit accrual, and support services, upon enrollment, without interruption in service provision. This recommendation is supported by the chapter 4 findings of

feature importance analysis, as well as was included in chapter 2 literature review under the findings on student mobility and disruption of instruction, student disengagement and, subsequently, dropout risk (Grigg, 2012; Rumberger, 2020). In addition to specific strategies, it is critical to monitor that stabilization supports are distributed equitably and do not ward others off transfers for another reason (e.g., school climate). Relatedly, districts should prioritize flagged schools as high-priority outreach schools and not as high-risk schools in a deterministic way.

Cushion Economic Strain Where Poverty Concentrates

Since low-income enrollment was a significant predictor of dropout risk, and supported by chapter 4 findings and chapter 2 literature (Rumberger, 2020), districts should enhance embedded supports in schools with the highest proportions of students from low-income backgrounds or schools with the worst outcomes in terms of dropping out of school. For example, low-income immersion in these high-poverty schools could result in additional access to overall assistance with basic needs, tutoring before or after school, structured after-school programming, and/or counseling and mentoring as part of an integrated school-based team. In addition to the findings produced from model technologies, the concentration of poverty is a consistent correlational to elevated dropout risk. Data systems are suggested for districts to examine saturation levels of supports and assess impacts on equity over time. Because this model is predictive (not causal), these examples are appropriate as risk mitigation strategies (not guaranteed improvements).

Maintain Disruption-Ready Continuity and Re-Engagement Plans

Given the significance of COVID-19 period variable, schools should remain engaged with plans for efficiently responding to instructional disruption due to a pandemic, infrastructure

collapse, or weather-related closure. Instructional continuity planning should include notifications and protocols for remote instruction, targeted counseling, re-entry supports, credit audits, and other strategies to re-engage students after significant time away from school. This recommendation is aligned with the scholarship on dropout prevention in the COVID-era, in addition to the temporal pattern observed in the model outputs (Bauman & Cranney, 2020; Rumberger, 2020; Seymour et al., 2020). Continuity should be viewed as student-driven and proactive, rather than disciplinary actions. It is also important to note that, while the model is not treated as causal, predictive evidence provides further support to this strategy as part of district-level risk preparation.

Strengthen Disability-Responsive Retention Supports

Although not the most influential feature at the final model, IEP enrollment was a prominent variable. Schools should ensure that disability-responsive planning is fully integrated into their dropout prevention planning. This could be through, for example, proactively contextualized transition planning, timely IEP reviews, case management, IEP-specific access to credit recovery options, and providing additional academic and behavioral support among others. This recommendation builds from the model evidence, as well as Chapter 2, that suggested students with disabilities have higher dropout risk (Irwin et al., 2021; McCauley, 2017). At the school level, justification for expanding services is provided by model-guided prioritization, and at the student level, research provides a little bit of guidance about what that service might be. The pattern of implementation should center on equity and family and staff input.

Operationalize Data-Informed Early Warning and Triage, With Guardrails

Given the predictive lift over the linear baseline with ensemble learners, districts should utilize the feature set from the study, and build a stakeholder-facing dashboard to flag schools or

cohorts for outreach. The whole pattern of implementation should be executed with an analytics pose, as laid out in the dissertation: business understanding, careful data preparation, comparative modeling, and out-of-sample evaluation to calibrate and make transparent alerts (Hayat Suhendar & Widyani, 2023). Districts ought to represent the outputs from the ensemble learners with signals potentially explaining features as opposed to simple baselines. It may also be beneficial for staff to become familiar with probabilities instead of absolutes. The on-track literature summarized in Chapter 2 strengthens the notion of early, intervention-based indicators, and is in tandem with the philosopher of using the outputs of models to prioritize supports, not shortcut educator judgment (Allensworth & Easton 2007). Considering the observations made in chapter 4, since the residuals were substantially more dispersed in the test set than the training set, districts should pilot the dashboards, locally validate thresholds, and refine cut-scores before scaling and diffusing the methodology throughout the system.

Treat Program-Adjacent Signals as Diagnostic, Not Prescriptive

The emergence of PE days as a significant predictor should be interpreted as an indicator for possible school engagement or climate; not be interpreted as an opportunity to effect change. As demonstrated by feature-importance profile in chapter 4, such indicators may covary with other conditions. Modifying one program variable cannot decrease the risk of dropout devoid of addressing more profound issues such as poverty and mobility. This is consistent literature in chapter 2 that dropout is caused by a variety of interacting forces (Rumberger, 2020). In practice, program-adjacent features should prompt diagnostic inquiry to finding out whether engagement opportunities are equitably accessible; whether adult-student relationships are supportive; and whether mixed-methods evidence, local data, model flags, and instructor insight should all inform which particular levers to prioritize.

Scope and Proportionality

The recommendations can only be applied to the concrete study design. The models were based on school level aggregated data (2017-2022), and the out-of-sample fit was moderate, with larger residual scattered across the test set, so anything derived from the results should be thought of as risk-aligned external supports and observed for impact and equity over time. Where the model is elevating high-leverage factors such as mobility, pandemic-period disruption, or poverty, there are credible literature examples to support those factors in Chapters 2. Where signals were less significance, like PE days, they should lead to inquiry, but not serve as a policy determiner, (Bauman & Cranney, 2020; Grigg, 2012; Rumberger, 2020; Seymour et al., 2020). This is a proportionate stance to the institutional expectations of making recommendations based on presented findings, reviewed within the literature, and not overstating findings in conclusion.

Recommendations for Future Research

This section offers specific next steps for researchers. These concrete recommendations come from the study's framework, findings, and limitations. They are still made from a predictive and not causal basis, based on modeling behavior and the limits of the determined data.

Evaluate Temporal Robustness and Extend Post-2022

The first recommendation is to evaluate the robustness of the model over time via time-aware validation and extending post-period. Residuals in chapter 4 depicted unbiased training predictions but wider error variation on the test data. Future research should train models on pre-COVID years and evaluate them on COVID/post-COVID years, and vice versa, to test performance drift and temporal stability.

As new data from the Illinois Report Card is released, expanding the panel of data beyond 2022 would permit formal checks of model decay, re-calibration, and relevance to the steady state. This falls within the evaluation stage of CRISP-DM and the survival-analysis frame employed in this study related to whether the COVID-period hazard shift holds over time. These steps would also directly address the current study's limitation related to random (non-temporal) splits.

Replicate in New Geographies and Within-State Subgroups

The second recommendation is to replicate the pipeline in other states or across demographically distinct regions within Illinois. The study acknowledged limitations in external validity; therefore, repeating the CRISP-DM pipeline across varied geographic or socioeconomic contexts can reveal whether feature rankings hold stable or are contingent on local conditions.

Stratified analysis within Illinois, for example, by district size, poverty level, or mobility intensity, can test whether key predictors behave similarly across subgroups. These replications would clarify the generalizability of findings and confirm whether *mobility*, *poverty*, and *pandemic-period disruption* consistently act as statewide risk signals or vary with setting.

Improve Data Completeness and Enable Individual-Level Models

The third recommendation is to improve construct measurement and expand the data infrastructure needed for deeper modeling. Chapter 3 documented removal of several variables with more than 50% missingness and pruning of collinear variables like *year* and *chronically truant*. While necessary for model stability, these decisions may obscure important relationships.

State and district agencies should enhance future predictive research by improving data completeness, promoting standardized reporting, and disaggregating subgroup indicators (e.g.,

race-specific dropout, attendance by program). Improved data fidelity would support re-inclusion of currently pruned variables and enable re-estimation of models with a fuller feature set.

In parallel, if student-level data become available, researchers should adopt multilevel (hierarchical) models that nest students within schools. These models would allow joint estimation of student-level risk factors (e.g., attendance, grades, IEP status) and school-level characteristics (e.g., mobility, poverty), with random intercepts or slopes and cross-level interaction terms. These refinements would address ecological fallacy concerns and expand the explanatory power of future predictive modeling in education.

The fourth recommendation is to bridge predictive and causal questions using designs appropriate for observational data while maintaining the non-experimental boundaries of the current study. Future research should begin with multivariable (covariate-adjusted) models to control for confounding and to test interactions among high-leverage predictors. Researchers should also consider panel or fixed-effects models (e.g., school and year effects), difference-in-differences for policy comparisons (e.g., pre-/post-COVID), and interrupted time series designs tied to implementation events.

Where possible, researchers may apply propensity score methods, doubly robust estimators, or instrumental variables if appropriate instruments or discontinuities exist. Natural experiments tied to funding formula changes, attendance reporting policy shifts, or service mandates provide opportunities to test these methods. Even in predictive frameworks, the interpretability of future studies can be improved by reporting the results of sensitivity analyses of training periods, feature sets, and weightings. While causality cannot be assumed from those analyses, they limit the opportunity to speculate in interpretation and contribute to the credibility of model-based narratives of the data. If warranted, in some cases, counterfactual scenarios can

be provided (e.g., how changes in dropout risk can be projected if mobility were reduced), while making clear that the focus is on describing an exploratory scenario rather than a treatment effect. These design choices respect the original study's predictive goal, but advance inferences about risk patterns beyond that predictive goal. This approach responds directly to the limitations noted in the methods chapter and aligns with the CRISP-DM emphasis on transparency and disciplined evaluation.

Advance Model Evaluation and Interpretability

The fifth recommendation is to strengthen model evaluation and interpretability in alignment with the CRISP-DM framework. This study paired feature-importance profiles with residual diagnostics and emphasized out-of-sample metrics. Future work should extend this by adding: (a) formal stability checks on importance rankings across validation folds, (b) comparisons between random, grouped, and time-aware splits, and (c) subgroup-level error analyses for equity-sensitive evaluation. Researchers can also test whether alternative ensemble structures or regularization paths reproduce the same ordering of influential predictors. This would increase confidence that *mobility*, *COVID-19 period*, and *low-income enrollment* are not just artifacts of a single model or sample.

Because the current study deliberately avoided opaque dimensionality reduction to preserve interpretability, future research should continue this practice. Principal components or similar methods may be included as supplementary analysis (e.g., appendices), but not as replacements for interpretable features. These refinements support both transparency and theoretical alignment, particularly with the study's survival-analysis framing that treats predictors as modifiers of a school-year dropout hazard.

Pilot a Real-World Early-Warning System

The sixth recommendation is to conduct a prospective pilot that operationalizes the predictive pipeline developed in this study. The pilot should use CRISP-DM phases to:

- Reconfirm business understanding with local stakeholders
- Refresh data preparation using the most current year
- Train and evaluate models under time-aware or grouped splits
- Track model performance and error behavior in real-time use

Because the unit of analysis is the school-year, pilots should assess outcomes at that level (e.g., accuracy of high-risk school flags, dropout rate trends across flagged and non-flagged groups). All model-informed decisions should include human-in-the-loop oversight, in line with the study's non-causal scope.

The pilot would determine if the $R^2 = 0.467$ achieved here translates to meaningful, actionable supportive decision-making. It would also provide metadata, such as differential variations by school size or type, which are necessary in case there is a need to refine models for execution in the future. It would also support taking the pipeline in real-world direction with the research-to-practice loop completed consistent with the analytic framework and methodological rigor established in this dissertation.

Replicate the Study Including Race and Ethnicity

The seventh recommendation is to replicate the study across demographically distinct regions within Illinois or in other states including race and ethnicity with the inclusion of student race and ethnicity variables. Ongoing statewide and national trends consistently demonstrate disparities in high school dropout rates across different student populations, and incorporating these demographic indicators would strengthen the analytical rigor and equity relevance of the

findings. While the present study was designed with the intent to include race and ethnicity data, limitations in data availability prevented their integration into the final analysis. Addressing this gap in subsequent studies would allow for more nuanced interpretation of differential outcomes and support more targeted, equity-informed interventions. These replications would address the current study's limitation related to race and ethnicity and clarify the generalizability of findings and confirm whether *mobility, poverty, and COVID-19 pandemic disruption* consistently as a statewide risk signals or vary with race and ethnicity.

Explore Different Data Science Techniques

The eighth recommendation is to explore the application of alternative and complementary data science techniques to model high school dropout outcomes. Methods such as logistic regression, among other classification approaches, may offer improved interpretability and predictive performance when examining binary outcomes like dropout status. Employing multiple modeling strategies would enhance robustness, enable comparative evaluation of model performance, and provide deeper insight into the relative influence of key predictors. These replications would clarify the generalizability of findings and confirm whether *mobility*, *poverty*, and *COVID-19 pandemic disruption* consistently emerge as top predictors across various modeling techniques.

Finally, it is important to specify that the term mobility, as used in this study, refers specifically to inter-school mobility, defined as student movement between schools during the academic period of analysis. This distinction is intentional and necessary to avoid confusion with economic or social mobility, which are conceptually and analytically distinct constructs. Clearly differentiating inter-school mobility ensures conceptual precision and supports accurate interpretation of the study's findings.

Conclusions

This dissertation addresses a persistent statewide challenge, elevated high school dropout rates in Illinois (2017–2022), by demonstrating how school-year–level predictive modeling can inform practical, risk-aligned decision support. Rather than explaining dropout in purely descriptive terms, the study shows that modern ensemble learners, applied within a transparent analytics workflow, can capture meaningful structure in statewide administrative data. The result is an actionable lens for screening and triage that complements educator judgment and local context.

Original Contributions to Educational Data Science. First, the study provides a replicable modeling pipeline that contrasts linear and ensemble approaches (multiple linear regression, random forest, XGBoost) under held-out evaluation, yielding XGBoost $R^2 \approx 0.467$ as a realistic benchmark for school-level prediction. Second, it integrates survival-analysis reasoning with machine learning by framing mobility, poverty concentration, and the COVID-19 period as risk-shifting exposures that alter the school-year hazard of dropout, thereby aligning theory with predictive evidence. Third, it codifies data governance choices (e.g., missingness thresholds, collinearity pruning, and non-causal interpretation of feature importance) that improve generalization and reproducibility and that are often under-specified in applied education analytics.

Synthesis of Problem, Method, and Findings. Using publicly available Illinois Report Card data and a CRISP-DM–guided process, the analysis established that student mobility, the COVID-19 period, and low-income enrollment are the most influential predictors of school-level dropout rates, with IEP enrollment and PE days providing additional signal. The ensembles' advantages over the linear baseline indicate that nonlinearities and interactions are consequential

in this domain, while the linear model remains valuable for transparent coefficient-based directionality. Taken together, these results justify treating predictions as probabilistic estimates suited to screening, prioritization, and planning, not as causal claims or individual-level determinations.

Broader Significance. The work offers a state-scale, policy-relevant template for early warning and resource allocation that is (a) technically rigorous enough to generalize beyond a single year, (b) interpretable enough for stakeholder communication, and (c) bounded by clear ethical and methodological guardrails (aggregate scope, non-causal use, human-in-the-loop). By elevating mobility stabilization, poverty-responsive supports, continuity planning, and disability-responsive retention as priority areas, the study connects predictive signal to concrete organizational levers without overreach. For researchers it provides a documented, reproducible baseline to extend with time-aware validation and richer covariates; for educators and policymakers it provides a decision-support framework to focus limited resources where the modeled risk is highest.

References

Adasme, P., Viveros, A., Ayub, M. S., Soto, I., Firoozabadi, A. D., & Rodríguez, D. Z. (2023, November). A multiple linear regression approach to optimize the worst-user capacity and power allocation in a wireless network. In *2023 South American Conference on Visible Light Communications (SACVLC)* (pp. 6–11). IEEE. https://doi.org/10.1109/SACVLC59022.2023.10347689

Ahmed, W., Wani, M. A., Plawiak, P., Meshoul, S., Mahmoud, A., & Hammad, M. (2025). Machine-learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions. *Scientific Reports, 15*(1), Article 12353. https://doi.org/10.1038/s41598-025-12353-4

Aina, C., Baici, E., Casalone, G., & Pastore, F. (2021). The determinants of university dropout: A review of the socio-economic literature. *Socio-Economic Planning Sciences, 77*, Article 101102. https://doi.org/10.1016/j.seps.2021.101102

Aldowah, H., Al-Samarraie, H., Alzahrani, A., & Alalwan, N. (2020). Factors affecting student dropout in MOOCs: A cause-and-effect decision-making model. *Journal of Computing in Higher Education, 32*, 429–454. https://doi.org/10.1007/s12528-019-09241-y

Al-Fakih, A., Al-Wajih, E., Saleh, R. A., & Muhit, I. B. (2024). Ensemble machine-learning models for predicting the CO_2 footprint of GGBFS-based geopolymer concrete. *Journal of Cleaner Production, 472*, Article 143463. https://doi.org/10.1016/j.jclepro.2024.143463

Aljohani, A., & Aburasain, R. Y. (2024). A hybrid framework for glaucoma detection through federated machine-learning and deep-learning models. *BMC Medical Informatics and Decision Making, 24*, Article 115. https://doi.org/10.1186/s12911-024-02518-y

Allensworth, E. M., & Easton, J. Q. (2007). *What matters for staying on-track and graduating in Chicago public high schools: A close look at course grades, failures, and attendance in the freshman year.* University of Chicago Consortium on School Research.

Alspaugh, J. W. (2000). The effect of transition grade to high school, gender, and grade level upon dropout rates. *American Secondary Education, 29*(1), 2–9.

André, A., Tessier, D., Louvet, B., & Girard, E. (2023). Teachers' perception of classes' engagement, observed motivating teaching practices, and students' motivation: A mediation analysis. *Social Psychology of Education: An International Journal.* https://doi.org/10.1007/s11218-023-09805-y

Annie E. Casey Foundation, & Fiester, L. (2010). *Early warning! Why reading by the end of third grade matters (KIDS COUNT Special Report).* Annie E. Casey Foundation.

Ansong, D., Okumu, M., Nyoni, T., Appiah-Kubi, J., Amoako, E. O., Koomson, I., & Conklin, J. (2023). The effectiveness of financial-capability and asset-building interventions in improving youth's educational well-being: A systematic review. *Adolescent Research Review.* https://doi.org/10.1007/s40894-023-00223-x

Avval, T. G., Moeini, B., Carver, V., Fairley, N., Smith, E. F., Baltrusaitis, J., Fernandez, V., Tyler, B. J., Gallagher, N., & Linford, M. R. (2021). The often-overlooked power of summary statistics in exploratory data analysis: Comparison of pattern-recognition entropy (PRE) to other summary statistics and introduction of divided spectrum-PRE (DS-PRE). *Journal of Chemical Information and Modeling, 61*(9), 4173–4189. https://doi.org/10.1021/acs.jcim.1c00244

Ayodele, O. (2023). *Exploratory data analysis with Python cookbook*. Packt Publishing.

Azad, C., Bhushan, B., Sharma, R., Shankar, A., Singh, K. K., & Khamparia, A. (2022). Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus. *Multimedia Systems, 28*(4), 1289–1307. https://doi.org/10.1007/s00530-021-00817-2

Balawi, M., & Tenekeci, G. (2024). Time-series traffic-collision analysis of London hotspots: Patterns, predictions and prevention strategies. *Heliyon, 10*(4), Article e25710. https://doi.org/10.1016/j.heliyon.2024.e25710

Balfanz, R., Herzog, L., & Iver, D. J. M. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist, 42*(4), 223–235. https://doi.org/10.1080/00461520701621079

Benden, D. K., & Lauermann, F. (2022). Students' motivational trajectories and academic success in math-intensive study programs: Why short-term motivational assessments matter. *Journal of Educational Psychology, 114*(5), 1062–1085. https://doi.org/10.1037/edu0000708

Berg, N., & Nelson, T. D. (2016). Pregnancy and dropout: Effects of family, neighborhood, and high-school characteristics on girls' fertility and dropout status. *Population Research and Policy Review, 35*(6), 757–789.

Bichri, H., Chergui, A., & Hain, M. (2024). Investigating the impact of train/test split ratio on the performance of pre-trained models with custom datasets. *International Journal of Advanced Computer Science & Applications, 15*(2), 331–339. https://doi.org/10.14569/ijacsa.2024.0150235

Booker, K., Sass, T. R., Gill, B., & Zimmer, R. (2010). The unknown world of charter high schools. *Education Next, 10*(2), 70–75.

Boualaphet, K., & Goto, H. (2020). Determinants of school dropout in Lao People's Democratic Republic: A survival analysis. *Journal of International Development, 32*(6), 961–975. https://doi.org/10.1002/jid.3486

Cassel, R. N. (2003). Use of personal development test to identify high-school and college-dropout students. *Education, 123*(4), Article 641.

Chan, J. Y., Leow, S. M., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z., & Chen, Y. (2021). Mitigating the multicollinearity problem and its machine-learning approach: A review. *Mathematics, 10*(8), Article 1283. https://doi.org/10.3390/math10081283

Chan, W. Y., Kuperminc, G. P., Seitz, S., Wilson, C., & Khatib, N. (2020). School-based group mentoring and academic outcomes in vulnerable high-school students. *Youth & Society, 52*(7), 1220–1237. https://doi.org/10.1177/0044118X19864834

Chen, C., Sonnert, G., Sadler, P. M., Sasselov, D. D., Fredericks, C., & Malan, D. J. (2020). Going over the cliff: MOOC dropout behavior at chapter transition. *Distance Education, 41*(1), 6–25. https://doi.org/10.1080/01587919.2020.1724772

Chen, H.-L., Chen, P., Zhang, Y., Xing, Y., Guan, Y.-Y., Cheng, D.-X., & Li, X.-W. (2020). Retention of volunteers and factors influencing program performance of the Senior-Care Volunteers Training Program in Jiangsu, China. *PLOS ONE, 15*(8), Article e0237390. https://doi.org/10.1371/journal.pone.0237390

Chen, Q., & Lee, S. (2021). A machine-learning approach to predict customer usage of a home workout platform. *Applied Sciences, 11*(21), Article 9927. https://doi.org/10.3390/app11219927

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression-analysis evaluation. *PeerJ Computer Science, 7*, Article e623. https://doi.org/10.7717/peerj-cs.623

Chitsamatanga, B. B., & Rembe, N. S. (2020). Children's rights to education in South Africa twenty years after democracy: A reflection on achievements, problems and areas for future action. *E-BANGI Journal, 17*(5), 99–118.

Churchill, E., Rogers, M., & Pristawa, K. (2021). High schoolers' and middle-schoolers' connections in their schools: Relation to tardiness, absences, disciplinary referrals, and failed courses. *National Youth Advocacy and Resilience Journal, 4*(2). https://doi.org/10.20429/nyarj.2021.040203

Chyn, E., & Katz, L. F. (2021). Neighborhoods matter: Assessing the evidence for place effects. *Journal of Economic Perspectives, 35*(4), 197–222.

Cobre, J., Tortorelli, F. A. C., & de Oliveira, S. C. (2019). Modeling two types of heterogeneity in the analysis of student success. *Journal of Applied Statistics, 46*(14), 2527–2539. https://doi.org/10.1080/02664763.2019.1601164

Cocoradă, E., Curtu, A. L., Năstasă, L. E., & Vorovencii, I. (2021). Dropout intention, motivation, and socio-demographics of forestry students in Romania. *Forests, 12*(5), Article 618. https://doi.org/10.3390/f12050618

Collett, D. (2023). *Modeling survival data in medical research* (3rd ed.). Chapman & Hall/CRC. https://doi.org/10.1201/9781003282525

Colpo, M. P., Primo, T. T., & Sanchotene de Aguiar, M. (2024). Lessons learned from the student-dropout patterns during the COVID-19 pandemic: An analysis supported by machine learning. *British Journal of Educational Technology, 55*(2), 560–585. https://doi.org/10.1111/bjet.13380

Conto, C., Akseer, S., Dreesen, T., Kamei, A., Mizunoya, S., & Rigole, A. (2021). Potential effects of COVID-19 school closures on foundational skills and country responses for mitigating learning loss. *International Journal of Educational Development, 87*, Article 102434. https://doi.org/10.1016/j.ijedudev.2021.102434

Costa, A. G., Mattos, J. C. B., Primo, T. T., Cechinel, C., & Munoz, R. (2021). Model for prediction of student dropout in a computer-science course. In *2021 XVI Latin American Conference on Learning Technologies (LACLO)* (pp. 137–143). IEEE. https://doi.org/10.1109/LACLO54177.2021.00020

Coussement, K., Phan, M., De Caigny, A., Benoit, D. F., & Raes, A. (2020). Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit-leaf model. *Decision Support Systems, 135*, Article 113325. https://doi.org/10.1016/j.dss.2020.113325

Crosnoe, R. (2009). Low-income students and the socioeconomic composition of public high schools. *American Sociological Review, 74*(5), 709–730. https://doi.org/10.1177/000312240907400502

Dass, S., Gary, K., & Cunningham, J. (2021). Predicting student dropout in self-paced MOOC courses using random forest models. *Information, 12*(11), Article 476. https://doi.org/10.3390/info12110476

Davis-Kean, P. E., Tighe, L. A., & Waters, N. E. (2021). The role of parent educational attainment in parenting and children's development. *Current Directions in Psychological Science, 30*(2), 186–192. https://doi.org/10.1177/0963721421993116

de la Cruz Orozco, I., & Heredia Rubio, B. (2019). Asistencia y deserción escolar de la juventud indígena en secundaria. *Revista Electrónica de Investigación Educativa, 21*(1), Article e24. https://doi.org/10.24320/redie.2019.21.e24.1973

Deleña, R. D., Dia, N. J., Sacayan, R. R., Sieras, J. C., Khalid, S. A., Macatotong, A. H. T., & Gulam, S. B. (2025). Predicting student retention: A comparative study of machine-learning approaches utilizing sociodemographic and academic factors. *Systems and Soft Computing, 7*, Article 200352. https://doi.org/10.1016/j.sasc.2025.200352

Dupéré, V., Dion, E., Nault-Brière, F., Archambault, I., Leventhal, T., & Lesage, A. (2018). Revisiting the link between depression symptoms and high-school dropout: Timing of exposure matters. *Journal of Adolescent Health, 62*(2), 205–211. https://doi.org/10.1016/j.jadohealth.2017.09.024

Emmerson, J., & Brown, J. M. (2021). Understanding survival analysis in clinical trials. *Clinical Oncology, 33*(1), 12–14.

Evans, J. S., Murphy, M. A., Holden, Z. A., & Cushman, S. A. (2010). Modeling species distribution and change using random forest. In *Predictive species and habitat modeling in landscape ecology: Concepts and applications* (pp. 139–159). Springer.

Farrugia, C., & Bhandari, R. (2020). Global trends in student mobility. In *The international encyclopedia of higher education systems and institutions* (pp. 560–568). Springer Netherlands.

Foreman-Murray, L., Krowka, S., & Majeika, C. E. (2022). A systematic review of the literature related to dropout for students with disabilities. *Preventing School Failure: Alternative Education for Children and Youth, 66*(3), 228–237. https://doi.org/10.1080/1045988X.2022.2037494

Gausel, N., & Bourguignon, D. (2020). Dropping out of school: Explaining how concerns for the family's social-image and self-image predict anger. *Frontiers in Psychology, 11*, Article 1868. https://doi.org/10.3389/fpsyg.2020.01868

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. *International Journal of Intelligent Technology and Applied Statistics, 11*(2), 105–111.

Ginevra, M. C., Di Maggio, I., Valbusa, I., Santilli, S., & Nota, L. (2021). Teachers' attitudes towards students with disabilities: The role of the type of information provided in the students' profiles of children with disabilities. *European Journal of Special Needs Education, 37*, 357–370.

Girgin, P., & Cabaroğlu, N. E. Ş. E. (2021). Web 2.0 Supported Flipped Learning Model: EFL Students' Perceptions and Motivation. *Cukurova University Faculty of Education Journal, 50*(2), 858-876. https://doi.org/10.14812/cuefd.944217

Goga, N., Radu, M. D., Vasilateanu, A., Dragomir, R., Buligan, R. M., Dinu, B. G., Moroti-Constantinescu, M.-I., Popvici, A. F., & Scurtu, D. (2021). ROSE – An intelligent system for student counseling and tutoring. In *2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)* (pp. 1–5). https://doi.org/10.1109/ECAI52376.2021.9515154

Gomez-Vazquez, A., Cruz-Tamayo, A.-A., Camacho-Perez, E., Chaves-Gurgel, A.-L., Herrera-Camacho, J., Mota-Rojas, D., García-Herrera, R.-A., Vinhas-Ítavo, L.-C., Dias-Silva, T.-P., & Chay-Canul, A.-J. (2024). Relationship between body weight and dorsal area in female buffaloes. *Revista Colombiana de Ciencias Pecuarias, 37*(4), 243–250. https://doi.org/10.17533/udea.rccp.v38n1a3

Grigg, J. (2012). School enrollment changes and student achievement growth: A case study in educational disruption and continuity. *Sociology of Education, 85*(4), 388–404. https://doi.org/10.1177/0038040712441374

Haeberlein, K., Handal, P. J., & Evans, L. (2021). Academic differences between an urban nativity school and an urban public school district. *Psychological Reports, 124*(6), 2703–2720. https://doi.org/10.1177/0033294120967274

Hagaman, J. L., & Casey, K. J. (2018). Teacher attrition in special education: Perspectives from the field. *Teacher Education and Special Education, 41*(4), 277–291.

Hamasha, M. M., Ali, H., Hamasha, S., & Ahmed, A. (2022). Ultra-fine transformation of data for normality. *Heliyon, 8*(5), Article e09370. https://doi.org/10.1016/j.heliyon.2022.e09370

Hardré, P. L., Davis, K. A., & Sullivan, D. W. (2008). Measuring teacher perceptions of the "how" and "why" of student motivation. *Educational Research and Evaluation, 14*(2), 155–179. https://doi.org/10.1080/13803610801956689

Hasani, I., & Kamberi, F. (2024). Empowering communities in Kosovo: The vital role of local government in advancing education and curbing deviant behavior. *Journal of Liberty and International Affairs, 10*(1).

Hayat Suhendar, M. T., & Widyani, Y. (2023). Machine learning application development guidelines using CRISP-DM and SCRUM concept. In *2023 IEEE International Conference on Data and Software Engineering (ICoDSE)* (pp. 168–173). https://doi.org/10.1109/ICoDSE59534.2023.10291438

Hazra, A., & Gogtay, N. (2017). Biostatistics series module 9: Survival analysis. *Indian Journal of Dermatology, 62*(3), 251–257. https://doi.org/10.4103/ijd.IJD_201_17

Hegde, H., Shimpi, N., Panny, A., Glurich, I., Christie, P., & Acharya, A. (2019). MICE vs PPCA: Missing data imputation in healthcare. *Informatics in Medicine Unlocked, 17.* https://doi.org/10.1016/j.imu.2019.100275

Hossain, M., Shamim Azad, S. B. M., Hossen, M. L., Khan, S. I., & Masum, A. K. M. (2022). Predictive analysis on university dropout rate of Bangladesh in COVID-19. In *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)* (pp. 439–444). https://doi.org/10.1109/ICISET54810.2022.9775898

Hussar, B., Zhang, J., Hein, S., Wang, K., Roberts, A., Cui, J., Smith, M., Mann, F. B., Barmer, A., Dilig, R., National Center for Education Statistics (ED), & American Institutes for Research (AIR). (2020). *The Condition of Education 2020* (NCES 2020-144). National Center for Education Statistics.

Iida, T. (2024). Identifying causes of errors between two wave-related data using performance metrics. *Applied Ocean Research, 148*, 104024. https://doi.org/10.1016/j.apor.2024.104024

Illinois State Board of Education. (2023). *Data and accountability report card data library*. https://www.isbe.net/ilreportcarddata.

Imola, C. P., & Krisztina, D. (2019). Dropping out of vocational training – Hungarian experiences. *Annals of the University of Oradea, Economic Science Series, 28*(2), 261–271.

In, J., & Lee, D. K. (2018). Survival analysis: Part I – analysis of time-to-event. *Korean Journal of Anesthesiology, 71*(3), 182–191. https://doi.org/10.4097/kja.d.18.00067

Inoue, T., Ichikawa, D., Ueno, T., Cheong, M., Inoue, T., Whetstone, W. D., Endo, T., Nizuma, K., & Tominaga, T. (2020). XGBoost, a machine-learning method, predicts neurological recovery in patients with cervical spinal-cord injury. *Neurotrauma Reports, 1*(1), 8. https://doi.org/10.1089/neur.2020.0009

Institute of Education Sciences. (2021). *State and district strategies to reduce dropouts*. https://ies.ed.gov/ncee/pubs/2021004/pdf/2021004.pdf

Irwin, V., Zhang, J., Wang, X., Hein, S., Wang, K., Roberts, A., York, C., Barmer, A., Bullock Mann, F., Dilig, R., Parker, S., Nachazel, T., Barnett, M., Purcell, S., National Center for Education Statistics (ED), & American Institutes for Research (AIR). (2021). *Report on the Condition of Education 2021* (NCES 2021-144). National Center for Education Statistics.

Jalota, S., & Suthar, M. (2024). Prediction of Marshall stability of asphalt concrete reinforced with polypropylene fiber using different soft-computing techniques. *Soft Computing – A Fusion of Foundations, Methodologies & Applications, 28*(2), 1425–1444. https://doi.org/10.1007/s00500-023-08339-x

Jimenez, O., Jesús, A., & Wong, L. (2023). Model for the prediction of dropout in higher education in Peru applying machine-learning algorithms: Random Forest, decision tree, neural network and support vector machine. In *2023 33rd Conference of Open Innovations Association (FRUCT)* (pp. 116–124). https://doi.org/10.23919/FRUCT58615.2023.10143068

Jin, C. (2023). MOOC student dropout prediction model based on learning behavior features and parameter optimization. *Interactive Learning Environments, 31*(2), 714–732. https://doi.org/10.1080/10494820.2020.1802300

Jinbo, Z., Yufu, L., & Haitao, M. (2025). Handling missing data using the XGBoost-based multiple-imputation-by-chained-equations regression method. *Frontiers in Artificial Intelligence, 8*, Article 1553220. https://doi.org/10.3389/frai.2025.1553220

Kabathova, J., & Drlik, M. (2021). Towards predicting student dropout in university courses using different machine-learning techniques. *Applied Sciences, 11*(7), Article 3130. https://doi.org/10.3390/app11073130

Kalnins, A., & Praitis-Hill, K. (2023). The VIF score: What is it good for? Absolutely nothing. *Organizational Research Methods, 28*(1), 58–75. https://doi.org/10.1177/10944281231216381

Kaneko, H. (2017). A new measure of regression-model accuracy that considers applicability domains. *Chemometrics and Intelligent Laboratory Systems, 171*, 1–8.

Kang, Y. S., & Chang, Y. J. (2019). Using a motion-controlled game to teach four elementary-school children with intellectual disabilities to improve hand hygiene. *Journal of Applied Research in Intellectual Disabilities, 32*(4), 942–951. https://doi.org/10.1111/jar.12587

Kearney, C. A., & Childs, J. (2023). Improving school-attendance data and defining problematic and chronic school absenteeism: The next stage for educational policies and health-based practices. *Preventing School Failure, 67*(4), 265–275. https://doi.org/10.1080/1045988X.2022.2124222

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine-learning approach. *European Journal of Higher Education, 10*(1), 28–47. https://doi.org/10.1080/21568235.2020.1718520

Kim, B.-Y., Cha, J. W., & Chang, K.-H. (2021). Twenty-four-hour cloud cover calculation using a ground-based imager with machine learning. *Atmospheric Measurement Techniques, 14*(10), 6695–6710. https://doi.org/10.5194/amt-14-6695-2021

Kloos, E., Nacik, E., & Ward, C. (2023). Developing implementation capacity of a state-education agency to improve outcomes for students with disabilities. *Journal of Disability Policy Studies, 34*(2), 127–136. https://doi.org/10.1177/10442073221096393

Koç, M., Zorbaz, O., & Demirtaş-Zorbaz, S. (2020). Has the ship sailed? The causes and consequences of school dropout from an ecological viewpoint. *Social Psychology of Education*, 23(5), 1149–1171. https://doi.org/10.1007/s11218-020-09568-w

Košir, S., Aslan, M., & Lakshminarayanan, R. (2023). Application of school-attachment factors as a strategy against school dropout: A case study of public-school students in Albania. *Children and Youth Services Review*, Article 107085.

Kroesch, A. M., & Peeples, K. N. (2021). High-school general-education teachers' perceptions of students with significant disabilities in inclusive settings. *Research, Advocacy, and Practice for Complex and Chronic Conditions, 40*(1), 26–41. https://doi.org/10.14434/rapcc.v40i1.31756

Kumar, G. K., Viswanath, P., & Rao, A. A. (2016). Ensemble of randomized soft decision trees for robust classification. *Sādhanā: Academy Proceedings in Engineering Sciences, 41*(3), 273–282.

Langsetmo, L., Schousboe, J. T., Taylor, B. C., Cauley, J. A., Howard A. Fink, H. A., Cawthon, P. M., Kado, D. M., Ensrud, K. E., & for the Osteoporotic Fractures in Men (MrOS) Research Group. (2023). Advantages and disadvantages of random-forest models for prediction of hip-fracture risk versus mortality risk in the oldest old. *JBMR Plus, 7*(8). https://doi.org/10.1002/jbm4.10757

LeBoeuf, W. A., & Fantuzzo, J. W. (2018). Effects of intradistrict school mobility and high student-turnover rates on early reading achievement. *Applied Developmental Science, 22*(1), 43–57. https://doi.org/10.1080/10888691.2016.1211481

Lee, K. G., & Polachek, S. W. (2018). Do school budgets matter? The effect of budget referenda on student dropout rates. *Education Economics, 26*(2), 129–144.

Lee-St. John, T. J., Walsh, M. E., Raczek, A. E., Vuilleumier, C. E., Foley, C., Heberle, A., Sibley, E., & Dearing, E. (2018). The long-term impact of systemic student support in elementary school: Reducing high-school dropout. *AERA Open, 4*(4). https://doi.org/10.1177/2332858418799085

Lequia, J. L., Vincent, L. B., Lyons, G. L., Asmus, J. M., & Carter, E. W. (2023). Individualized education programs of high-school students with significant disabilities. *Education and Training in Autism and Developmental Disabilities, 58*(1), 22–35. https://doi.org/10.1177/215416472305800103

Li, I., & Carroll, D. (2020). Factors influencing dropout and academic performance: An Australian higher-education equity perspective. *Journal of Higher Education Policy and Management, 42*, 14–30. https://doi.org/10.1080/1360080X.2019.1649993

Lichand, G., Doria, C. A., Leal-Neto, O., & Fernandes, J. P. C. (2022). The impacts of remote learning in secondary education during the pandemic in Brazil. *Nature Human Behaviour, 6*(8), 1079–1086. https://doi.org/10.1038/s41562-022-01350-6

Lickteig, S. J., & Lickteig, A. (2019). Forgotten and overlooked: A personal reflection of foster parenting and school. *Educational Considerations, 44*(2). https://doi.org/10.4148/0146-9282.2178

Liu, L., Chen, Z., Wang, Y., & Liu, G. (2022). Predicting gasoline RON loss by machine learning. In *2022 IEEE International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC)* (pp. 21–267). https://doi.org/10.1109/SDPC55702.2022.9915813

Lorenzo-Lledó, A., Lorenzo, G., Lledó, A., & Pérez-Vázquez, E. (2020). Inclusive methodologies from the teaching perspective for improving performance in university students with disabilities. *Journal of Technology and Science Education, 10*, 127–141. https://doi.org/10.3926/jotse.887

Manoj, R., Abhishek, S., Nair, B. P. Anjali, T., & Ramlal, N. P. (2023). A contemporary method of assessing water quality based on the fusion of predictive analytics and deep-structured learning. In *2023 8th International Conference on Communication and Electronics Systems* (pp. 961–967). https://doi.org/10.1109/ICCES57224.2023.10192729

Martínez-Carrascal, J. A., Hlosta, M., & Sancho-Vinuesa, T. (2023). Using survival analysis to identify populations of learners at risk of withdrawal: Conceptualization and impact of demographics. *International Review of Research in Open and Distributed Learning, 24*(1). https://doi.org/10.19173/irrodl.v24i1.6589

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., HernandezOrallo, J., Kull, M., Lachiche, N. J. A. H., Ramírez-Quintana, M. J., & Flach, P. A. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*. Advance online publication. https://doi.org/10.1109/TKDE.2019.2962680

Mau, W. C. J., & Li, J. (2018). Factors influencing STEM-career aspirations of underrepresented high-school students. *The Career Development Quarterly, 66*(3), 246–258. https://doi.org/10.1002/cdq.12146

McCauley, E. J. (2017). The cumulative probability of arrest by age 28 years in the United States by disability status, race/ethnicity, and gender. *American Journal of Public Health, 107*(12), 1977–1981. https://doi.org/10.2105/AJPH.2017.304095

McFarland, J., Cui, J., Holmes, J., Wang, X., National Center for Education Statistics (ED), & American Institutes for Research (AIR). (2020). *Trends in high-school dropout and completion rates in the United States: 2019. Compendium report* (NCES 2020-117). National Center for Education Statistics.

McFarland, J., Cui, J., Rathbun, A., & Holmes, J. (2018). *Trends in high-school dropout and completion rates in the United States: 2018. Compendium report* (NCES 2019-117). National Center for Education Statistics.

Melo, E. C., & de Souza, F. S. H. (2023). Improving the prediction of school dropout with the support of the semi-supervised learning approach. *iSys–Brazilian Journal of Information Systems, 16*(1), 10-1. https://doi.org/10.5753/isys.2023.2852

Meng, H., Lu, Y., Tian, Z., Jiang, X., Han, Z., & Niu, J. (2023). Performance-evaluation method of day-ahead load-prediction models in a district-heating and cooling system: A case study. *Energies, 16*(14), Article 5402. https://doi.org/10.3390/en16145402

Miller, T. (2011). *Partnering for education reform: Remarks by the U.S. Deputy Secretary, Tony Miller at the Church of God in Christ's International 51 AIM Convention in Houston, TX.* Washington, DC: U.S. Department of Education. http://www.ed.gov/news/speeches/partneringeducation-reform

Mohanasundaram, V., & Rangaswamy, B. (2025). Elastic net with Bayesian-density-estimation model for feature selection for photovoltaic-energy prediction. *Scientific Reports, 15*(1), 1–23. https://doi.org/10.1038/s41598-025-92633-1

Montes, G. C., & Mendes, L. (2021). Effects of violence on school dropout: A panel-data analysis for Rio de Janeiro. *Journal of Developing Areas, 55*(4), Article 329.

Mtey, A. R. (2024). Community involvement in education provision for indigenous-pastoral-community girls in Tanzania. *Interchange: A Quarterly Review of Education, 55*(1), 27–49. https://doi.org/10.1007/s10780-024-09511-4

Mughal, A. W. (2020). Secondary-school students who drop out of school in rural Pakistan: The perspectives of fathers. *Educational Research, 62*(2), 199–215. https://doi.org/10.1080/00131881.2020.1755604

Munk, T., Rui, N., Zhu, W., & Carlson, E. (2021). Using state-data sets and meta-analysis of low-powered studies to evaluate a school-based dropout-prevention program for students with disabilities. *Studies in Educational Evaluation, 68*. https://doi.org/10.1016/j.stueduc.2020.100969

Myers, J., Witzel, B., Bouck, E., & Mathis, J. (2021). Middle-school math teachers' perceptions of their classroom practices among students with disabilities before and during the pandemic: A pilot study. *Journal of Online Learning Research, 7*(3), 209–231.

Narvaez, J. L., & Gomez, E. (2023). Factors associated with school dropout in a public educational institution in Cartagena (Colombia). *Panorama Económico, 31*(3), 265–288. https://doi.org/10.32997/pe-2023-4707

National Center for Education Statistics. (2023). *Status dropout rates. Condition of Education.* U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/programs/coe/indicator/coj

National Center for Education Statistics. (2024). *Status dropout rates. Condition of Education.* U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/programs/coe/indicator/coj

Neild, R. C. (2009). Falling off track during the transition to high school: What we know and what can be done. *The Future of Children, 19*(1), 53–76. https://doi.org/10.1353/foc.0.0020

Newburger, E., Correll, M., & Elmqvist, N. (2023). Fitting Bell curves to data distributions using visualization. *IEEE Transactions on Visualization and Computer Graphics, 29*(12), 5372–5383. https://doi.org/10.1109/TVCG.2022.3210763

Nicho, M., Hamed, A., Gaber, T., & Al Arimi, J. H. (2025). A-XGBoost: A resilient machine-learning technique for predicting crimes against women across cultures on low-cardinality crime data. *Cogent Social Sciences, 11*(1). https://doi.org/10.1080/23311886.2025.2527392

Novosel, L. M. (2022). Understanding the evidence: Quantitative research designs. *Urologic Nursing, 42*(6), 303–311. https://doi.org/10.7257/2168-4626.2022.42.6.303

Nygren, T., Kronlid, D. O., Larsson, E., Novak, J., Bentrovato, D., Wasserman, J., Welply, O., Guath, M., & Anamika. (2020). Global citizenship education for global citizenship? Students' views on learning about, thought, and for human rights, peace, and sustainable development in England, India, New Zealand, South Africa, and Sweden. *Journal of Social Science Education, 19*(4), 63–97.

Orlova, E. V. (2021). Data-driven design to credit-risk management using digital-footprint intelligence. In *2021 3rd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)* (pp. 161–166). https://doi.org/10.1109/SUMMA53307.2021.9632188

Paksi, B., Széll, K., & Fehérvári, A. (2023). Empirical testing of a multidimensional model of school-dropout risk. *Social Sciences, 12*(2), 50. https://doi.org/10.3390/socsci12020050

Palma, M., Distefano, V., Giungato, G., & Mazuruse, G. (2025). Predicting odor concentration for environmental sustainability: A comparison among machine Learning methods. *Quality & Quantity*. https://doi.org/10.1007/s11135-025-02056-3

Pan, F., Huang, B., Zhang, C., Zhu, X., Wu, Z., Zhang, M., & Li, Z. (2022). A survival-analysis based volatility and sparsity-modeling network for student-dropout prediction. *PLOS ONE, 17*(5), Article e0267138. https://doi.org/10.1371/journal.pone.0267138

Papadakaki, M., Maraki, A., Bitsakos, N., & Chliaoutakis, J. (2022). Perceived knowledge and attitudes of faculty members towards inclusive education for students with disabilities: Evidence from a Greek university. *International Journal of Environmental Research and Public Health, 19*(4). https://doi.org/10.3390/ijerph19042151

Peset, F., Garzón-Farinós, F., González, L., García-Massó, X., Ferrer-Sapena, A., Toca-Herrera, J., & Sánchez-Pérez, E. (2020). Survival analysis of author keywords: An application to the library and information sciences area. *Journal of the Association for Information Science and Technology, 71*(4), 462–473. https://doi.org/10.1002/asi.24248

Pov, S., Kawai, N., & Murakami, R. (2020). Identifying causes of lower-secondary school dropout in Cambodia: A two-level hierarchical-linear model. *International Journal of Inclusive Education, 26*, 834–847. https://doi.org/10.1080/13603116.2020.1735542

Prins, H., Dörre, A., & Schmidt, D. (2023). Statutory health insurance-covered pre-exposure prophylaxis in Germany: Changing trends in nationwide tenofovir disoproxil/emtricitabine prescriptions during the COVID-19 pandemic. *Frontiers in Pharmacology*, 1–10. https://doi.org/10.3389/fphar.2023.1241310

Punyangarm, V., & Chotayakul, S. (2025). Hybrid sequence learning with interpretability for multi-class quality prediction in injection molding. *Results in Engineering, 27*, Article 106408. https://doi.org/10.1016/j.rineng.2025.106408

Rahmany, M., Zin, A. M., & Sundararajan, E. A. (2020). Comparing tools provided by Python and R for exploratory data analysis. *IJISCS (International Journal of Information System and Computer Science), 4*(3), 131–142. https://doi.org/10.56327/ijiscs.v4i3.933

Rai, S., Mishra, P., & Ghoshal, U. C. (2021). Survival analysis: A primer for the clinician-scientist. *Indian Journal of Gastroenterology, 40*(5), 541–549. https://doi.org/10.1007/s12664-021-01232-1

Ramsdal, G. H., & Wynn, R. (2022). Theoretical basis for a group intervention aimed at preventing high-school dropout: The case of "Guttas Campus." *International Journal of Environmental Research and Public Health, 19*(24). https://doi.org/10.3390/ijerph192417025

Reeves, S. (2021). Literacy and educational opportunities in China and the United States. *Delta Kappa Gamma Bulletin, 87*(5), 33–37.

Rekha, I. S., Shetty, J., & Basri, S. (2023). Students' continuance intention to use MOOCs: Empirical evidence from India. *Education and Information Technologies, 28*(4), 4265–4286. https://doi.org/10.1007/s10639-022-11308-w

Ressa, T., & Andrews, A. (2022). High-school dropout dilemma in America and the importance of reformation of education systems to empower all students. *International Journal of Modern Education Studies, 6*(2), 423–447. https://doi.org/10.51383/ijonmes.2022.234

Ribeiro, R., Pilastri, A., Moura, C., Rodrigues, F., Rocha, R., & Cortez, P. (2020). Predicting the tear strength of woven fabrics via automated machine-learning: An application of the CRISP-DM methodology. https://doi.org/10.5220/0009411205480555

Rose, R. A., Hopson, L. M., Bowen, G. L., & Glennie, E. (2012). Students' perceived parental school-behaviour expectations and their academic performance: A longitudinal analysis. *Family Relations, 61*, 175–191.

Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., & Scheinost, D. (2024). Data-leakage inflates prediction performance in connectome-based machine-learning models. *Nature Communications, 15*(1), 1–15. https://doi.org/10.1038/s41467-024-46150-w

Roure, C., Lentillon-Kaestner, V., & Pasco, D. (2021). Students' individual interest in physical education: Development and validation of a questionnaire. *Scandinavian Journal of Psychology, 62*(1), 64–73. https://doi.org/10.1111/sjop.12669

Rumberger, R. W. (2020). The economics of high-school dropouts. *The Economics of Education, 149-158.* https://doi.org/10.1016/B978-0-12-815391-8.00012-4

Saltz, J. S., & Shamshurin, I. (2016). Big data team process methodologies: A literature review and the identification of key factors for a project's success. In *2016 IEEE International Conference on Big Data (Big Data)* (pp. 2872–2879). https://doi.org/10.1109/BigData.2016.7840936

Sameer, S. K. L., & Sriramya, P. (2021). Improving the efficiency by novel feature-extraction technique using decision-tree algorithm compared with SVM classifier algorithm for predicting heart disease. *Alinteri Journal of Agriculture Sciences, 36*(1), 713–720. https://doi.org/10.47059/alinteri/V36I1/AJAS21100

Scarano, A., Rella Riccardi, M., Mauriello, F., D'Agostino, C., Pasquino, N., & Montella, A. (2023). Injury-severity prediction of cyclist crashes using random forests and random-parameters logit models. *Accident Analysis & Prevention, 192*, Article 107275. https://doi.org/10.1016/j.aap.2023.107275

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying the CRISP-DM process model. *Procedia Computer Science, 181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Segura, M., Mello, J., & Hernández, A. (2022). Machine-learning prediction of university-student dropout: Does preference play a key role? *Mathematics, 10*(18), Article 3359. https://doi.org/10.3390/math10183359

Seymour, K., Skattebol, J., & Pook, B. (2020). Compounding education disengagement: COVID-19 lockdown, the digital divide and wrap-around services. *Journal of Children's Services, 15*(4), 243–251. https://doi.org/10.1108/JCS-08-2020-0049

Shuja, A., Ali, A., Khan, S. S. A., Burki, S. B., & Bilal, S. (2022). Perspectives on the factors affecting students' dropout rate during COVID-19: A case study from Pakistan. *Sage Open*, *12*(2). https://doi.org/10.1177/21582440221097378

Snyder, T. D., de Brey, C., & Dillow, S. A. (2018). *Digest of Education Statistics, 2018* (54th ed.). National Center for Education Statistics. https://nces.ed.gov/pubs2020/2020009.pdf

Song, Z., Sung, S.-H., Park, D.-M., & Park, B.-K. (2023). All-year dropout prediction modeling and analysis for university students. *Applied Sciences*, *13*(2), Article 1143. https://doi.org/10.3390/app13021143

South, S. J., Haynie, D. L., & Bose, S. (2007). Student mobility and school dropout. *Social Science Research, 36*(1), 68–94. https://doi.org/10.1016/j.ssresearch.2005.10.001

Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N. A., Trollor, J., & Brodaty, H. (2020). A comparison of machine-learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports, 10*(1), Article 20410. https://doi.org/10.1038/s41598-020-77220-w

Subho, M. R., & Chowdhury, M. R. (2022). Performance evaluation of fake-news detection using predictive modeling. In *2022 25th International Conference on Computer and Information Technology (ICCIT)* (pp. 49–54). https://doi.org/10.1109/ICCIT.2022.9245457

Szlyk, H. S. (2020). Resilience among students at risk of dropout: Expanding perspectives on youth suicidality in a non-clinical setting. *School Mental Health, 12*(3), 567–579. https://doi.org/10.1007/s12310-020-09366-…

Talatahari, S., Chen, F., & Gandomi, A. H. (2025). Developing a robust machine-learning framework for predicting the behavior of large-scale structure. *Journal of Building Engineering, 105*, Article 112204. https://doi.org/10.1016/j.jobe.2025.112204

Tarwidi, D., Pudjaprasetya, S. R., Adytia, D., & Apri, M. (2022). An optimized XGBoost-based machine-learning method for predicting wave run-up on a sloping beach. *MethodsX, 10*, Article 102119. https://doi.org/10.1016/j.mex.2023.102119

Tomaszewski, W., Zajac, T., Rudling, E., te Riele, K., McDaid, L., & Western, M. (2023). Uneven impacts of COVID-19 on the attendance rates of secondary-school students from different socioeconomic backgrounds in Australia: A quasi-experimental analysis of administrative data. *Australian Journal of Social Issues, 58*(1), 111–130. https://doi.org/10.1002/ajs4.219

Uldall, J. S., & Rojas, C. G. (2022). An application of machine-learning in public policy: Early-warning prediction of school dropout in the Chilean public-education system. *Multidisciplinary Business Review.*

U.S. Bureau of Labor Statistics. (2021). *Earnings and unemployment rates by educational attainment.* https://www.bls.gov/emp/chart-unemployment-earnings-education.htm

U.S. Department of Education. (2022). *Special Education-Grants to States.* https://www2.ed.gov/programs/osepgts/index.html

van Den Berghe, L., Vandevelde, S., & De Pauw, S. S. W. (2022). School dropout as the result of a complex interplay between individual and environmental factors: A study on the perspectives of support workers. *Teachers & Teaching, 28*(5), 603–617. https://doi.org/10.1080/13540602.2022.2062748

Voulgarides, C. K., & Barrio, B. L. (2021). The Individuals with Disabilities Education Act (IDEA) and the equity imperative: Examining early-childhood transitions to special education. *Multiple Voices, 21*(1), 40–54.

Vrigazova, B. (2021). The proportion for splitting data into training and test sets for the bootstrap in classification problems. *Business Systems Research, 12*(1), 228–242. https://doi.org/10.2478/bsrj-2021-0015

Wabba, M. A., & House, R. J. (1974). Expectancy theory in work and motivation: Some logical and methodological issues. *Human Relations, 27*(2), 121–147. https://doi.org/10.1177/001872677402700202

Wan, Y. (2022). Capacity or money? Why students choose to drop out of junior high school in rural northeast China. *Educational Review, 74*(7), 1264–1281. https://doi.org/10.1080/00131911.2021.1887818

Wen, Y., Tian, Y., Wen, B., Zhou, Q., Cai, G., & Liu, S. (2020). Consideration of the local correlation of learning behaviors to predict dropouts from MOOCs. *Tsinghua Science and Technology.* https://doi.org/10.26599/TST.2019.9010013

Xing, W., & Du, D. (2019). Dropout prediction in MOOCs: Using deep learning for personalized intervention. *Journal of Educational Computing Research, 57*, 547–570. https://doi.org/10.1177/0735633118757015

Yılmaz, A. B., & Karataş, S. (2022). Why do open and distance-education students drop out? Views from various stakeholders. *International Journal of Educational Technology in Higher Education, 19*(1), 1–22. https://doi.org/10.1186/s41239-022-00333-x

Zeng, X. (2022, September). Survival-analysis techniques for censored data. In *2022 International Conference on Applied Physics and Computing (ICAPC)* (pp. 70–73). IEEE.

Zhao, Y., & Li, H. (2022). Research on data-analysis method based on multiple linear-regression model and grey model. In *2022 10th International Conference on Information Systems and Computing Technology (ISCTech)* (pp. 324–327). https://doi.org/10.1109/ISCTech58360.2022.00057

Appendix A

Description of Variables in the Dataset

Variable Name		Description	Type of data	Level of measurement
County		It refers to an administrative division in the state of Illinois.	Categorical	Nominal
District type		It refers to the type of school district in the state of Illinois. A school district is organized and established to provide instruction up to and including grade 12.	Categorical	Nominal
School type		It refers to the type of school in a school district.	Categorical	Nominal
District size		It refers to the size (small, medium, and large) of a school district in the state of Illinois.	Categorical	Ordinal
Grades served		It refers to the grade level of education that a school and/or district provides for general education. This study focuses on grades 9 through 12, which are grade levels taught in high school.	Categorical	Nominal

Number of student enrollment		It refers to the total student enrollment in the school and district as of October 1 of the school year.	Numerical	Ratio
Rate of student enrollment white		It refers to the proportion of white students enrolled in the school and district as of October 1 of the school year.	Numerical	Ratio
Rate of student enrollment black or African American		It refers to the proportion of black or African American students enrolled in the school and district as of October 1 of the school year.	Numerical	Ratio
Rate of student enrollment Hispanic or Latino		It refers to the proportion of Hispanic or Latino students enrolled in the school and district as of October 1 of the school year.	Numerical	Ratio
Rate of student Enrollment Asian		It refers to the proportion of Asian students enrolled in the school and district as of October 1 of the school year.	Numerical	Ratio
Rate of student enrollment with disabilities		It refers to the proportion of students with disabilities enrolled in the school and district as of October 1 of the school year.	Numerical	Ratio

Rate of student enrollment English learners		It refers to the proportion of English learners enrolled in the school and district as of October 1 of the school year.	Numerical	Ratio
Rate of student enrollment with low-income		It refers to the proportion of students from low-income families enrolled in the school and district as of October 1 of the school year.	Numerical	Ratio
Rate of student enrollment homeless		It refers to the proportion of homeless students enrolled in the school and district as of October 1 of the school year.	Numerical	Ratio
Student attendance rate		It refers to the weighted measure of the number of days a student is present relative to the total number of potential attendance days.	Numerical	Ratio
Student mobility rate		It refers to the unduplicated count for students who transferred in and out of the serving school at any time during the school year.	Numerical	Ratio
Student mobility rate English learners		It refers to the proportion of English learners students who transferred in and out of the serving school at any time	Numerical	Ratio

		during the school year.		
Student mobility rate with IEP		It refers to the proportion of students with IEP who transferred in and out of the serving school at any time during the school year.	Numerical	Ratio
Chronically truant students		It refers to the number of students subject to compulsory attendance who have been absent without valid cause from such attendance for 5 percent or more of the previous 180 regular attendance days.	Numerical	Ratio
Student chronic truancy rate		It refers to the proportion of students subject to compulsory attendance who have been absent without valid cause from such attendance for 5 percent or more of the previous 180 regular attendance days.	Numerical	Ratio
High school dropout rate total		It refers to the proportion of high school dropouts in an entity (school, district, state) per enrollment. Dropouts are students who are of school age but are no longer	Numerical	Ratio

		enrolled or did not graduate from high school.		
High school dropout rate male		The proportion of male students who are of school age but are no longer enrolled or did not graduate from high school.	Numerical	Ratio
High school dropout rate females		The proportion of female students who are of school age but are no longer enrolled or did not graduate from high school.	Numerical	Ratio
High school dropout rate white		The proportion of white students who are of school age but are no longer enrolled or did not graduate from high school.	Numerical	Ratio
High school dropout rate black or African American		The proportion of black or African American students who are of school age but are no longer enrolled or did not graduate from high school.	Numerical	Ratio
High school dropout rate Hispanic or Latino		The proportion of Hispanic or Latino students who are of school age but are no longer enrolled or did not graduate from high school.	Numerical	Ratio
High school dropout rate Asian		The proportion of Asian students who are of school age but are no longer enrolled or	Numerical	Ratio

		did not graduate from high school.		
High school dropout rate English learners		The proportion of English learners students who are of school age but are no longer enrolled or did not graduate from high school.	Numerical	Ratio
High school dropout with IEP		The proportion of students with IEP who are of school age but are no longer enrolled or did not graduate from high school.	Numerical	Ratio
High school dropout rate with low-income		The proportion of students from low-income families who are of school age but are no longer enrolled or did not graduate from high school.	Numerical	Ratio
Average class size high school		It refers to the average number of students in each class in a high school as of the last day of school.	Numerical	Ratio
Average number of days of physical education per week per student		It refers to the average number of days a student engages in a course of physical education in a week.	Numerical	Ratio
Teacher retention rate		It refers to the proportion of full-time teachers in the same school/district in the past three years.	Numerical	Ratio

Principal turnover within six years		It refers to the number of different principals at the same high school in the last six years.	Numerical	Ratio

Appendix B

Codebook

This appendix provides the codes used in the analysis. These codes were not edited so they can be copied and used for reproducibility purposes. The following codes were used to perform the multiple linear regression, random forest, and XGboost models to conduct this study.

```
# Necessary libraries
from sklearn.dummy import DummyRegressor
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error
from sklearn.model_selection import RandomizedSearchCV
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.model_selection import train_test_split
from statsmodels.stats.outliers_influence import variance_inflation_factor
import numpy as np
from sklearn.preprocessing import PowerTransformer
from scipy.stats import skew
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import LabelEncoder
from scipy.stats.mstats import winsorize
```

```
from scipy.stats import zscore
from scipy.stats import shapiro
from scipy import stats
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.linear_model import Ridge, Lasso, ElasticNet
from sklearn.model_selection import GridSearchCV
from sklearn.pipeline import Pipeline
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.linear_model import ElasticNetCV
# Install and Import XGBoost
!pip install xgboost
from xgboost import XGBRegressor
#  Import the Data into the session (From Google Drive)
from google.colab import drive
drive.mount('/content/drive')
import pandas as pd
import numpy as np
# Load raw datasets
df2017 = pd.read_excel("/content/drive/MyDrive/ReducedDatasets/ReducedDataset2017.xlsx")
df2018 = pd.read_excel("/content/drive/MyDrive/ReducedDatasets/ReducedDataset2018.xlsx")
df2019 = pd.read_excel("/content/drive/MyDrive/ReducedDatasets/ReducedDataset2019.xlsx")
df2020 = pd.read_excel("/content/drive/MyDrive/ReducedDatasets/ReducedDataset2020.xlsx")
df2021 = pd.read_excel("/content/drive/MyDrive/ReducedDatasets/ReducedDataset2021.xlsx")
```

```
df2022 = pd.read_excel("/content/drive/MyDrive/ReducedDatasets/ReducedDataset2022.xlsx")
Renaming dictionaries
rename_2017 = {
    "County": "County",
    "District Type ": "District_Type",
    "District Size ": "District_Size",
    "School Type": "School_Type",
    "Grades Served": "Grades_Served",
    "Student Enrollment - Total": "Enrolment_Total",
    "Student Enrollment - White %": "Enrolment_White",
    "Student Enrollment - Black or African American %": "Enrolment_Black",
    "Student Enrollment - Hispanic or Latino %": "Enrolment_Hispanic",
    "Student Enrollment - Asian %": "Enrolment_Asian",
    "Student Enrollment - EL %": "Enrolment_EL",
    "Student Enrollment - IEP %": "Enrolment_IEP",
    "Student Enrollment - Low Income %": "Enrolment_Low_Income",
    "Student Enrollment - Homeless %": "Enrolment_Homeless",
    "Student Attendance Rate": "Attendance_Rate",
    "Student Mobility Rate": "Mobility_Rate",
    "Student Mobility Rate - EL": "Mobility_Rate_EL",
    "Student Mobility Rate - IEP": "Mobility_Rate_IEP",
    "Chronically Truant Students": "Chronically_Truant",
    "Student Chronic Truancy Rate": "Chronic_Truancy_Rate",
    "High School Dropout Rate - Total": "Dropout_Rate_Total",
    "High School Dropout Rate - Male": "Dropout_Rate_Male",
    "High School Dropout Rate - Female": "Dropout_Rate_Female",
    "High School Dropout Rate - White": "Dropout_Rate_White",
    "High School Dropout Rate - Black or African American": "Dropout_Rate_Black",
```

```
    "High School Dropout Rate - Hispanic or Latino": "Dropout_Rate_Hispanic",
    "High School Dropout Rate - Asian": "Dropout_Rate_Asian",
    "High School Dropout Rate - EL": "Dropout_Rate_EL",
    "High School Dropout Rate - IEP": "Dropout_Rate_IEP",
    "High School Dropout Rate - Low Income": "Dropout_Rate_Low_Income",
    "Avg Class Size - High School": "Class_Size_High",
    "Avg Number of Days of Physical Education per Week Per Student": "PE_Days",
    "Teacher Retention Rate": "Retention_Rate",
    "Principal Turnover within 6 Years": "Turnover"
}
rename_2018 = {
    "County": "County",
    "District Type": "District_Type",
    "District Size": "District_Size",
    "School Type": "School_Type",
    "Grades Served": "Grades_Served",
    "Student Enrollment - Total": "Enrolment_Total",
    "Student Enrollment - White %": "Enrolment_White",
    "Student Enrollment - Black or African American %": "Enrolment_Black",
    "Student Enrollment - Hispanic or Latino %": "Enrolment_Hispanic",
    "Student Enrollment - Asian %": "Enrolment_Asian",
    "Student Enrollment - EL %": "Enrolment_EL",
    "Student Enrollment - IEP %": "Enrolment_IEP",
    "Student Enrollment - Low Income %": "Enrolment_Low_Income",
    "Student Enrollment - Homeless %": "Enrolment_Homeless",
    "Student Attendance Rate": "Attendance_Rate",
    "Student Mobility Rate": "Mobility_Rate",
    "Student Mobility Rate - EL": "Mobility_Rate_EL",
```

```
    "Student Mobility Rate - IEP": "Mobility_Rate_IEP",
    "Chronically Truant Students": "Chronically_Truant",
    "Student Chronic Truancy Rate": "Chronic_Truancy_Rate",
    "High School Dropout Rate - Total": "Dropout_Rate_Total",
    "High School Dropout Rate - Male": "Dropout_Rate_Male",
    "High School Dropout Rate - Female": "Dropout_Rate_Female",
    "High School Dropout Rate - White": "Dropout_Rate_White",
    "High School Dropout Rate - Black or African American": "Dropout_Rate_Black",
    "High School Dropout Rate - Hispanic or Latino": "Dropout_Rate_Hispanic",
    "High School Dropout Rate - Asian": "Dropout_Rate_Asian",
    "High School Dropout Rate - EL": "Dropout_Rate_EL",
    "High School Dropout Rate - IEP": "Dropout_Rate_IEP",
    "High School Dropout Rate - Low Income": "Dropout_Rate_Low_Income",
    "Avg Class Size - High School": "Class_Size_High",
    "Avg Number of days of Physical Education Per Week Per Student": "PE_Days",
    "Teacher Retention Rate": "Retention_Rate",
    "Principal Turnover within 6 Years": "Turnover"
}
rename_2019 = {
    "County": "County",
    "District Type": "District_Type",
    "District Size": "District_Size",
    "School Type": "School_Type",
    "Grades Served": "Grades_Served",
    "# Student Enrollment": "Enrolment_Total",
    "% Student Enrollment - White": "Enrolment_White",
    "% Student Enrollment - Black or African American": "Enrolment_Black",
    "% Student Enrollment - Hispanic or Latino": "Enrolment_Hispanic",
```

```
    "% Student Enrollment - Asian": "Enrolment_Asian",
    "% Student Enrollment - IEP": "Enrolment_IEP",
    "% Student Enrollment - EL": "Enrolment_EL",
    "% Student Enrollment - Low Income": "Enrolment_Low_Income",
    "% Student Enrollment - Homeless": "Enrolment_Homeless",
    "Student Attendance Rate": "Attendance_Rate",
    "Student Mobility Rate": "Mobility_Rate",
    "Student Mobility Rate - EL": "Mobility_Rate_EL",
    "Student Mobility Rate - IEP": "Mobility_Rate_IEP",
    "Chronically Truant Students": "Chronically_Truant",
    "Student Chronic Truancy Rate": "Chronic_Truancy_Rate",
    "High School Dropout Rate - Total": "Dropout_Rate_Total",
    "High School Dropout Rate - Male": "Dropout_Rate_Male",
    "High School Dropout Rate - Female": "Dropout_Rate_Female",
    "High School Dropout Rate - White": "Dropout_Rate_White",
    "High School Dropout Rate - Black or African American": "Dropout_Rate_Black",
    "High School Dropout Rate - Hispanic or Latino": "Dropout_Rate_Hispanic",
    "High School Dropout Rate - Asian": "Dropout_Rate_Asian",
    "High School Dropout Rate - EL": "Dropout_Rate_EL",
    "High School Dropout Rate - IEP": "Dropout_Rate_IEP",
    "High School Dropout Rate - Low Income": "Dropout_Rate_Low_Income",
    "Avg Class Size - High School": "Class_Size_High",
    "Avg Number of Days of Physical Education per Week Per Student": "PE_Days",
    "Teacher Retention Rate": "Retention_Rate",
    "Principal Turnover within 6 Years": "Turnover"
}
2020, 2021, and 2022 reuse the same pattern as 2019
rename_2020 = rename_2019.copy()
```

```
rename_2021 = rename_2019.copy()
rename_2022 = rename_2019.copy()
# Standardized column order
standard_columns = [
    "County", "District_Type", "District_Size", "School_Type", "Grades_Served",
    "Enrolment_Total", "Enrolment_White", "Enrolment_Black", "Enrolment_Hispanic",
    "Enrolment_Asian", "Enrolment_IEP", "Enrolment_EL", "Enrolment_Low_Income",
    "Enrolment_Homeless", "Attendance_Rate", "Mobility_Rate", "Mobility_Rate_EL",
    "Mobility_Rate_IEP", "Chronically_Truant", "Chronic_Truancy_Rate",
    "Dropout_Rate_Total", "Dropout_Rate_Male", "Dropout_Rate_Female",
    "Dropout_Rate_White", "Dropout_Rate_Black", "Dropout_Rate_Hispanic",
    "Dropout_Rate_Asian", "Dropout_Rate_EL", "Dropout_Rate_IEP",
    "Dropout_Rate_Low_Income", "Class_Size_High", "PE_Days",
    "Retention_Rate", "Turnover"
]
# Cleaning helper function
def clean_dataset(df, rename_dict, year):
    # Rename columns
    df = df.rename(columns=rename_dict)
    # Ensure ALL standardized columns exist (add missing ones as NaN)
    for col in standard_columns:
        if col not in df.columns:
            df[col] = None
    # Select in correct order
    df = df[standard_columns]
    # Add year column
    df["Year"] = year
    return df
```

```
# Apply cleaning to each year
df2017_clean = clean_dataset(df2017, rename_2017, 2017)
df2018_clean = clean_dataset(df2018, rename_2018, 2018)
df2019_clean = clean_dataset(df2019, rename_2019, 2019)
df2020_clean = clean_dataset(df2020, rename_2020, 2020)
df2021_clean = clean_dataset(df2021, rename_2021, 2021)
df2022_clean = clean_dataset(df2022, rename_2022, 2022)
# Merge all datasets
df = pd.concat([
    df2017_clean,
    df2018_clean,
    df2019_clean,
    df2020_clean,
    df2021_clean,
    df2022_clean
], ignore_index=True)
# Preview final merged dataset
print(df.head())
print(df.shape)
#  Basic data transformation and cleaning
df.columns = df.columns.str.strip()     # Strip whitespace
df.columns = df.columns.str.lower()  # Makes them all lowercase
# remove surrounding spaces and make title case for the following string variables
start_col = 'county'
end_col = 'year'
# Get list of columns between start and end (inclusive)
cols = df.loc[:, start_col:end_col].columns
# Clean each column: strip spaces and make title case
```

```
for col in cols:
    df[col] = df[col].astype(str).str.strip().str.title()
#   Keep only rows where school_type is "High School"
final_df = df[df["school_type"] == "High School"]
#   Reset index after filtering
final_df = final_df.reset_index(drop=True)
#   Check if only high schools remain
print(final_df["school_type"].value_counts())
print(final_df["year"].value_counts())
Try converting any object column to numeric if possible
for col in final_df.select_dtypes(include='object').columns:
    converted = pd.to_numeric(final_df[col], errors='coerce')
    # Only overwrite if at least one value was converted successfully
    if converted.notna().sum() > 0:
        final_df[col] = converted
        print(f"Column '{col}' successfully converted to numeric.")
    else:
        print(f"Column '{col}' could not be converted to numeric.")
# Convert district size into numeric categories
# Check how the unique variables are stored
print(final_df["district_size"].value_counts())
  # Create a mapping dictionary
size_mapping = {
    'Small': 1,
    'Medium': 2,
    'Large': 3
}
Apply the mapping (NaNs are preserved by default)
```

```
final_df['district_size_encoded'] = final_df['district_size'].map(size_mapping)
   # Confirm if worked.
print(final_df["district_size_encoded"].value_counts())
#   Describe the data, then sum the missing values per column and get info for the datatypes
#   Count missing values per variable
print(final_df.isnull().sum())
#   Data types for the variables
print(final_df.info())
#    Those with more than 50% missing data be dropped
#  Create a heatmap to visualize the missing values
   # Set up the figure
plt.figure(figsize=(12, 8))
   # Create heatmap of missing values
sns.heatmap(df.isnull(),
            cbar=False,
            cmap='viridis',  # Options: 'viridis', 'magma', 'coolwarm', etc.
            yticklabels=False)
plt.title('Heatmap of Missing Values')
plt.xlabel('Variables')
plt.ylabel('Observations')
plt.tight_layout()
plt.show()
missing_threshold = 0.5  # 50% threshold
#   Calculate percentage of missing values per column
missing_percent = final_df.isnull().mean()
#   Filter columns where missing percentage is greater than 50%
high_missing_cols = missing_percent[missing_percent > missing_threshold]
```

```
#   Display the results
print("Columns with more than 5% missing values:")
print(high_missing_cols)
#   Drop these Columns
final_df.drop(columns=['grades_served', 'mobility_rate_el', 'dropout_rate_asian',
'enrolment_black', 'enrolment_hispanic','enrolment_white', 'enrolment_el', 'enrolment_homeless',
'mobility_rate_iep', 'dropout_rate_white','dropout_rate_black', 'dropout_rate_hispanic',
'dropout_rate_el','dropout_rate_female', 'dropout_rate_iep',
'dropout_rate_low_income','dropout_rate_male', 'enrolment_asian'], inplace=True)
#   Check final data without the columns
print(final_df.isnull().sum())


sns.heatmap(final_df.isnull(),
        cbar=False,
        cmap='viridis',  # Options: 'viridis', 'magma', 'coolwarm', etc.
        yticklabels=False)
plt.title('Heatmap of Missing Values')
plt.xlabel('Variables')
plt.ylabel('Observations')
plt.tight_layout()
plt.show()
#  Data Imputation
# Select only numeric columns
numeric_cols = final_df.select_dtypes(include=[np.number]).columns
# Perform Shapiro-Wilk test and store results
shapiro_results = []
for col in numeric_cols:
    stat, p_value = shapiro(final_df[col].dropna())  # Drop NaNs for test
```

```
    shapiro_results.append(
        {"Variable": col, "Test Statistic": stat, "p-value": p_value})
# Convert results to DataFrame
shapiro_df = pd.DataFrame(shapiro_results)
# Display the table
print(shapiro_df)
# Impute missing values with median for all numeric columns
final_df[numeric_cols] = final_df[numeric_cols].apply(
    lambda col: col.fillna(col.median()))
# Impute missing values in 'district_size' using the most frequent category (mode)
# Get the most common category
mode_value = final_df["district_size_encoded"].mode()[0]
final_df["district_size_encoded"] = final_df["district_size_encoded"].fillna(
    mode_value)  # Fill missing values
#   Check if missing values remain
print(final_df.isnull().sum())  # Should print 0 for numeric columns
print(final_df["year"].value_counts())
print(final_df["district_size_encoded"].value_counts())
# Descriptive statistics for numeric variables
def numeric_summary(final_df):
    summary = pd.DataFrame({
        'n': final_df.count(),
        'Min': final_df.min(),
        'Max': final_df.max(),
        'Mean': final_df.mean(),
        'SD': final_df.std(),
        'Median': final_df.median()
```

```
    })
    return summary.round(2)
# Apply function to numeric columns only
numeric_table = numeric_summary(final_df[numeric_cols])
print(numeric_table)
# Create histograms for each numeric column
numeric_cols = final_df.select_dtypes(include=[np.number]).columns
for col in numeric_cols:
    plt.figure(figsize=(8, 4))
    plt.hist(final_df[col].dropna(), bins=30, edgecolor='black')
    plt.title(f'Histogram of {col}')
    plt.xlabel(col)
    plt.ylabel('Frequency')
    plt.grid(True)
    plt.tight_layout()
    plt.show()
#    Check for outliers
#   Compute Z-scores for numeric columns
numeric_cols = final_df.select_dtypes(include=[np.number]).columns
z_scores = final_df[numeric_cols].apply(zscore)
#   Flag outliers where |Z| > 3
outliers_z = (z_scores.abs() > 3)
print(outliers_z.sum())  # Count of outliers per column
#    Label Encoding
#   Covid Period.
    # Initialize LabelEncoder
le = LabelEncoder()
```

```
    # Categorize and encode the covid_period variable in one step
final_df["covid_period"] = le.fit_transform(final_df["year"].apply(
    lambda x: "Pre-COVID" if x < 2020 else "During-COVID"))
# Check encoding
print(final_df["year"].value_counts())
print(final_df["covid_period"].value_counts())
print(final_df.info())
#   Scale the DataFrame
#   Selecting only numeric columns for scaling
numeric_cols = final_df.select_dtypes(include=[np.number]).columns
#   Standardizing
scaler = StandardScaler()
final_df[numeric_cols] = scaler.fit_transform(final_df[numeric_cols])
#   Generate Graphs
#   Histogram
final_df[numeric_cols].hist(figsize=(12, 10), bins=30)
plt.suptitle("Histograms of Numeric Variables")
plt.show()
#   Box Plots
plt.figure(figsize=(15, 8))
sns.boxplot(data=final_df[numeric_cols])
plt.xticks(rotation=90)
plt.title("Boxplots of Numeric Variables")
plt.show()
#   Apply Transformations Based on Skewness
# Compute skewness for each numeric column
skew_vals = final_df[numeric_cols].apply(skew).sort_values(ascending=False)
```

```
print(skew_vals)
#   Yeo-Johnson Transformation (For Both Right and Left Skewness)
pt = PowerTransformer(method='yeo-johnson')  # Works with negative values too
pt = PowerTransformer(method='yeo-johnson')  # Works with negative values too
final_df[numeric_cols] = pt.fit_transform(final_df[numeric_cols])

#   And check skewness again
print(final_df[numeric_cols].apply(skew).sort_values(ascending=False))
#   Another set of Histograms
final_df[numeric_cols].hist(figsize=(12, 10), bins=30)
plt.suptitle("Histograms After Transformation")
plt.show()
#  Compute correlation matrix
# Select only numeric columns
numeric_df = final_df.select_dtypes(include=[np.number])
# Compute correlation matrix
corr_matrix = numeric_df.corr()
# Display correlation values
print(corr_matrix)
#   Visualize the Correlation Matrix (Heatmap)
plt.figure(figsize=(12, 8))
sns.heatmap(corr_matrix, annot=True, fmt=".2f",
            cmap="coolwarm", linewidths=0.5)
plt.title("Correlation Matrix Heatmap")
plt.show()
#  Feature Engineering
# Compute VIF for each feature
```

```
numeric_df = numeric_df.dropna()  # Ensure no missing values
vif_data = pd.DataFrame()
vif_data["Feature"] = numeric_df.columns
vif_data["VIF"] = [variance_inflation_factor(
    numeric_df.values, i) for i in range(len(numeric_df.columns))]
print(vif_data.sort_values(by="VIF", ascending=False))
#   Drop these 2 since have high VIF
numeric_df.drop(columns=["chronically_truant","year"], inplace=True)
# Compute VIF for each feature again
numeric_df = numeric_df.dropna()  # Ensure no missing values
vif_data = pd.DataFrame()
vif_data["Feature"] = numeric_df.columns
vif_data["VIF"] = [variance_inflation_factor(
    numeric_df.values, i) for i in range(len(numeric_df.columns))]
print(vif_data.sort_values(by="VIF", ascending=False))
#      Predictive Modeling
# Define the target variable
target = "dropout_rate_total"
# Select all numeric features except the target variable
X = numeric_df.select_dtypes(include=[np.number]).drop(
    columns=[target], errors="ignore")
# Define the dependent variable
y = numeric_df[target]
# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42)
# Train Multiple Linear Regression Model
```

```
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)
y_pred_lr = lr_model.predict(X_test)
# Train Random Forest Model
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)
# Train the XGBoost Model
# Initialize XGBoost model
xgb_model = XGBRegressor(
    n_estimators=300,
    learning_rate=0.05,
    max_depth=6,
    subsample=0.8,
    colsample_bytree=0.8,
    random_state=42
)
# Fit the model
xgb_model.fit(X_train, y_train)
# Predictions
y_pred_xgb = xgb_model.predict(X_test)
# Evaluation Function (Stores Metrics in Variables)
def evaluate_model(y_true, y_pred, model_name):
    r2 = r2_score(y_true, y_pred)
    mae = mean_absolute_error(y_true, y_pred)
    rmse = np.sqrt(mean_squared_error(y_true, y_pred))
    print(f"\nModel: {model_name}")
```

```
    print(f"R² Score: {r2:.4f}")
    print(f"Mean Absolute Error (MAE): {mae:.4f}")
    print(f"Root Mean Squared Error (RMSE): {rmse:.4f}")
    return r2, mae, rmse
# Store metrics for later use
r2_lr, mae_lr, rmse_lr = evaluate_model(y_test, y_pred_lr, "Multiple Linear Regression")
r2_rf, mae_rf, rmse_rf = evaluate_model(y_test, y_pred_rf, "Random Forest")
r2_xgb, mae_xgb, rmse_xgb = evaluate_model(y_test, y_pred_xgb, "XGBoost")
#      Get Feature Importance
# Random Forest.
# Extract feature importance
feature_importances = rf_model.feature_importances_
# Create a DataFrame to display importance
importance_df = pd.DataFrame(
    {'Feature': X_train.columns, 'Importance': feature_importances})
# Sort by importance
importance_df = importance_df.sort_values(by='Importance', ascending=False)
# Plot feature importance
plt.figure(figsize=(10, 6))
plt.barh(importance_df['Feature'],
         importance_df['Importance'], color='skyblue')
plt.xlabel("Feature Importance")
plt.ylabel("Features")
plt.title("Random Forest Feature Importance")
plt.gca().invert_yaxis()  # Invert y-axis for better visualization
plt.show()
# Display top features
```

```
print(importance_df.head(10))  # Show top 10 most important features
# XGBoost
# Extract and plot feature importance
importance = pd.DataFrame({
    'Feature': X_train.columns,
    'Importance': xgb_model.feature_importances_
}).sort_values(by='Importance', ascending=False)
plt.figure(figsize=(10, 6))
sns.barplot(x='Importance', y='Feature', data=importance)
plt.title("Feature Importance - XGBoost")
plt.show()
#  Hyperparameter Tuning for linear regression, Random Forest, and XGBoost
# Multicollinearity Assessment
# 1️⃣ Calculate VIF (Before Elastic Net)
def calculate_vif(X):
    vif_data = pd.DataFrame()
    vif_data["Feature"] = X.columns
    vif_data["VIF"] = [variance_inflation_factor(X.values, i)
                       for i in range(X.shape[1])]
    return vif_data
vif_before = calculate_vif(X_train)
print("\n 🔹 Variance Inflation Factor (Before Elastic Net):")
print(vif_before.sort_values(by="VIF", ascending=False))
# 2️⃣ Elastic Net Regression for Feature Selection
elastic_cv = ElasticNetCV(l1_ratio=[0.1, 0.5, 0.7, 0.9, 1],
                          alphas=np.logspace(-3, 2, 50),
```

```
                cv=5,
                random_state=42)
elastic_cv.fit(X_train, y_train)
elastic_coeff = pd.DataFrame({
    "Feature": X_train.columns,
    "Coefficient": elastic_cv.coef_
}).sort_values(by="Coefficient", ascending=False)
print("\n 🔹 Elastic Net Coefficients (Feature Selection Results):")
print(elastic_coeff)
# Identify features retained by Elastic Net
selected_features = elastic_coeff[elastic_coeff["Coefficient"] != 0]["Feature"].tolist()
print("\n✅ Features Selected by Elastic Net:", selected_features)
# 3️⃣ Calculate VIF After Feature Selection
if len(selected_features) > 1:
    vif_after = calculate_vif(X_train[selected_features])
    print("\n 🔹 Variance Inflation Factor (After Elastic Net):")
    print(vif_after.sort_values(by="VIF", ascending=False))
else:
    print("\n⚠️ Only one feature selected, VIF not applicable.")
# Linear regression
# Standardize data + fit Ridge regression
ridge_pipeline = Pipeline([
    ('scaler', StandardScaler()),
    ('ridge', Ridge())
])
```

```
# Define parameter grid
param_grid = {
    'ridge__alpha': [0.01, 0.1, 1, 10, 100]
}
# Grid search with cross-validation
grid_search = GridSearchCV(ridge_pipeline, param_grid, cv=5, scoring='r2')
grid_search.fit(X_train, y_train)
# Best model and hyperparameters
best_lr = grid_search.best_estimator_
y_pred_lr = best_lr.predict(X_test)
print("Best R² score:", grid_search.best_score_)
print("Best alpha:", grid_search.best_params_['ridge__alpha'])
# Random Forest
# Define hyperparameter grid
param_dist = {
    'n_estimators': [50, 100, 200, 300, 500],  # Number of trees
    'max_depth': [10, 20, 30, None],  # Maximum depth of trees
    'min_samples_split': [2, 5, 10],  # Minimum number of samples to split
    'min_samples_leaf': [1, 2, 4],  # Minimum number of samples at a leaf
    'max_features': ['sqrt', 'log2']  # Number of features per split
}
# Initialize model
rf = RandomForestRegressor(random_state=42)
# Perform Randomized Search
rf_random = RandomizedSearchCV(
    estimator=rf,
    param_distributions=param_dist,
```

```
    n_iter=50,  # Number of iterations
    cv=5,  # Cross-validation folds
    scoring='r2',  # Optimize for R-squared
    verbose=2,
    n_jobs=-1,  # Use all processors
    random_state=42
)
# Fit the model
rf_random.fit(X_train, y_train)
# Best parameters
best_rf = rf_random.best_estimator_
y_pred_rf = best_rf.predict(X_test)
print("Best Hyperparameters:", rf_random.best_params_)
# Train the final model with best parameters
best_rf = RandomForestRegressor(
    n_estimators=500,
    min_samples_split=5,
    min_samples_leaf=4,
    max_features='sqrt',
    max_depth=10,
    random_state=42
)
# Fit the model on training data
best_rf.fit(X_train, y_train)

# Make predictions
y_pred = best_rf.predict(X_test)
```

```
# Evaluate model performance
r2 = r2_score(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
print(f"Tuned Model Performance:")
print(f"R² Score: {r2:.4f}")
print(f"Mean Absolute Error (MAE): {mae:.4f}")
print(f"Root Mean Squared Error (RMSE): {rmse:.4f}")
# XGBoost
param_grid_xgb = {
    'n_estimators': [100, 200, 300, 500],
    'max_depth': [3, 5, 7, 10],
    'learning_rate': [0.01, 0.05, 0.1, 0.2],
    'subsample': [0.6, 0.8, 1.0],
    'colsample_bytree': [0.6, 0.8, 1.0]
}
xgb_random = RandomizedSearchCV(
    estimator=XGBRegressor(random_state=42),
    param_distributions=param_grid_xgb,
    n_iter=30,
    cv=5,
    scoring='r2',
    verbose=2,
    n_jobs=-1,
    random_state=42
)
xgb_random.fit(X_train, y_train)
```

```
# Best parameters and model
best_xgb = xgb_random.best_estimator_
y_pred_best_xgb = best_xgb.predict(X_test)
print("Best Hyperparameters:", xgb_random.best_params_)
evaluate_model(y_test, y_pred_best_xgb, "XGBoost (Tuned)")
# Evaluation & Parsimony Metrics
def evaluate_model(y_true, y_pred, model_name, model_obj):
    r2 = r2_score(y_true, y_pred)
    mae = mean_absolute_error(y_true, y_pred)
    rmse = np.sqrt(mean_squared_error(y_true, y_pred))
    n_params = count_model_parameters(model_obj)
    print(f"\nModel: {model_name}")
    # Use a try-except block to handle models that don't have get_params
    try:
        print(f"Best Params: {model_obj.get_params()}")
    except AttributeError:
        print("Best Params: N/A (Model does not support get_params)")
    print(f"R² Score: {r2:.4f}")
    print(f"MAE: {mae:.4f}")
    print(f"RMSE: {rmse:.4f}")
    print(f"Parsimony (Fewer is better): {n_params} Parameters")
    return r2, mae, rmse, n_params
def count_model_parameters(model):
    if isinstance(model, Pipeline):
        # Corrected indentation
        model = model.named_steps['ridge'] # Assuming 'ridge' is the name in the pipeline
    if hasattr(model, 'coef_'):
```

```
        return np.count_nonzero(model.coef_)
    elif isinstance(model, RandomForestRegressor):
        # Sum the number of nodes in all trees
        total_nodes = 0
        if hasattr(model, 'estimators_'): # Check if estimators exist (after fitting)
          for est in model.estimators_:
              total_nodes += est.tree_.node_count
          return total_nodes
        else:
            return 0 # Model not yet fitted
    else:
        return 0  # Fallback
evaluate_model(y_test, y_pred_lr, "Linear Regression (Ridge)", best_lr)
evaluate_model(y_test, y_pred_rf, "Random Forest", best_rf)
evaluate_model(y_test, y_pred_best_xgb, "XGBoost (Tuned)", best_xgb)
#  Get feature importances
# Random Forest.
feature_importance = pd.DataFrame({
    'Feature': X_train.columns,
    'Importance': best_rf.feature_importances_
})
# Sort by importance
feature_importance = feature_importance.sort_values(
    by='Importance', ascending=False)
# Plot feature importance
plt.figure(figsize=(10, 6))
sns.barplot(x='Importance', y='Feature',
```

```
        data=feature_importance, hue='Feature', dodge=False, palette='viridis', legend=False)
plt.xlabel('Feature Importance Score')
plt.ylabel('Features')
plt.title('Feature Importance in Random Forest Model')
plt.show()
 # XGBoost
# Extract and plot feature importance
importance = pd.DataFrame({
    'Feature': X_train.columns,
    'Importance': best_xgb.feature_importances_
}).sort_values(by='Importance', ascending=False)
plt.figure(figsize=(10, 6))
sns.barplot(x='Importance', y='Feature', data=importance)
plt.title("Feature Importance - XGBoost")
plt.show()
#  Residual Analysis for the Random Forest Model and XGBoost
# Random Forest
# Predict on training data
y_train_pred = best_rf.predict(X_train)
y_test_pred = best_rf.predict(X_test)
# Calculate residuals
train_residuals = y_train - y_train_pred
test_residuals = y_test - y_test_pred
# Plot residuals
plt.figure(figsize=(12, 5))
# Training residuals
plt.subplot(1, 2, 1)
```

```
sns.histplot(train_residuals, kde=True, bins=30, color='blue')
plt.axvline(0, color='red', linestyle='dashed')
plt.title('Training Residuals Distribution')
# Test residuals
plt.subplot(1, 2, 2)
sns.histplot(test_residuals, kde=True, bins=30, color='green')
plt.axvline(0, color='red', linestyle='dashed')
plt.title('Test Residuals Distribution')
plt.show()
#   Residuals vs. Predicted Values
plt.figure(figsize=(10, 5))
sns.scatterplot(x=y_test_pred, y=test_residuals, alpha=0.7)
plt.axhline(0, color='red', linestyle='dashed')
plt.xlabel('Predicted Dropout Rate')
plt.ylabel('Residuals')
plt.title('Residuals vs. Predicted Values')
plt.show()
# XGBoost
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.datasets import make_regression
from xgboost import XGBRegressor
#  Generate synthetic regression data
X, y = make_regression(n_samples=500, n_features=5, noise=15, random_state=42)
#  Split into train and test sets
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# 3. Train an XGBoost model
xgb_model = XGBRegressor(
    n_estimators=200,
    learning_rate=0.1,
    max_depth=4,
    random_state=42,
    n_jobs=-1
)
xgb_model.fit(X_train, y_train)
# 4. Predict on train and test data
y_train_pred = xgb_model.predict(X_train)
y_test_pred = xgb_model.predict(X_test)
#  Calculate residuals
train_residuals = y_train - y_train_pred
test_residuals = y_test - y_test_pred
#  Plot residual histograms
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
sns.histplot(train_residuals, kde=True, bins=30, color='blue')
plt.axvline(0, color='red', linestyle='dashed')
plt.title('Training Residuals Distribution (XGBoost)')
plt.subplot(1, 2, 2)
sns.histplot(test_residuals, kde=True, bins=30, color='green')
plt.axvline(0, color='red', linestyle='dashed')
plt.title('Test Residuals Distribution (XGBoost)')
plt.show()
```

```
#  Residuals vs Predicted Values plot
plt.figure(figsize=(10, 5))
sns.scatterplot(x=y_test_pred, y=test_residuals, alpha=0.7)
plt.axhline(0, color='red', linestyle='dashed')
plt.xlabel('Predicted Values (XGBoost)')
plt.ylabel('Residuals')
plt.title('Residuals vs. Predicted Values (XGBoost)')
plt.show()
#  Baseline Models for Comparison
#   Baseline 1: Mean Predictor(Dummy Regressor)
# Initialize Dummy Regressor (Mean Predictor)
dummy_model = DummyRegressor(strategy="mean")
dummy_model.fit(X_train, y_train)
# Predictions
y_pred_dummy = dummy_model.predict(X_test)
# Evaluate Performance
r2_dummy = r2_score(y_test, y_pred_dummy)
mae_dummy = mean_absolute_error(y_test, y_pred_dummy)
rmse_dummy = np.sqrt(mean_squared_error(y_test, y_pred_dummy))
# Print results
print(f"Baseline (Mean Predictor) R² Score: {r2_dummy:.4f}")
print(f"Baseline (Mean Predictor) MAE: {mae_dummy:.4f}")
print(f"Baseline (Mean Predictor) RMSE: {rmse_dummy:.4f}")
## Baseline 2: Single Feature Linear Regression (mobility_rate)
# Use the properly scaled final_df directly
target = "dropout_rate_total"
y = numeric_df[target]
```

```
X_mobility = numeric_df[['mobility_rate']]  # Single feature
# Redo the train-test split
X_train, X_test, y_train, y_test = train_test_split(
    X_mobility, y, test_size=0.2, random_state=42)
baseline_lr = LinearRegression()
baseline_lr.fit(X_train, y_train)
y_pred_baseline_lr = baseline_lr.predict(X_test)
print("\n Baseline 2: Single Feature Linear Regression (mobility_rate)")
r2_baseline_lr = r2_score(y_test, y_pred_baseline_lr)
mae_baseline_lr = mean_absolute_error(y_test, y_pred_baseline_lr)
rmse_baseline_lr = np.sqrt(mean_squared_error(y_test, y_pred_baseline_lr))
print(f"R²: {r2_baseline_lr:.4f}")
print(f"MAE: {mae_baseline_lr:.4f}")
print(f"RMSE: {rmse_baseline_lr:.4f}")
# Store for comparison
r2_baseline_lr = r2_score(y_test, y_pred_baseline_lr)
mae_baseline_lr = mean_absolute_error(y_test, y_pred_baseline_lr)
rmse_baseline_lr = np.sqrt(mean_squared_error(y_test, y_pred_baseline_lr))
# Re-calculate evaluation metrics for Linear Regression, Random Forest, and XGBoost
# These were originally calculated in cell CmiIAPw9ts7Y
# Train Multiple Linear Regression Model
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)
y_pred_lr = lr_model.predict(X_test)
r2_lr = r2_score(y_test, y_pred_lr)
mae_lr = mean_absolute_error(y_test, y_pred_lr)
rmse_lr = np.sqrt(mean_squared_error(y_test, y_pred_lr))
```

```
# Train Random Forest Model
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)
r2_rf = r2_score(y_test, y_pred_rf)
mae_rf = mean_absolute_error(y_test, y_pred_rf)
rmse_rf = np.sqrt(mean_squared_error(y_test, y_pred_rf))
# Train the XGBoost Model
xgb_model = XGBRegressor(
    n_estimators=300,
    learning_rate=0.05,
    max_depth=6,
    subsample=0.8,
    colsample_bytree=0.8,
    random_state=42
)
xgb_model.fit(X_train, y_train)
y_pred_xgb = xgb_model.predict(X_test)
r2_xgb = r2_score(y_test, y_pred_xgb)
mae_xgb = mean_absolute_error(y_test, y_pred_xgb)
rmse_xgb = np.sqrt(mean_squared_error(y_test, y_pred_xgb))
#   Compare with Random Forest, Multiple Linear Regression, and XGBoost
# Store results in a DataFrame
results = pd.DataFrame({
    "Model": ["Baseline (Mean Predictor)", "Baseline (Single Feature LR)", "Multiple Linear
Regression", "Random Forest", "XGBoost"],
    "R² Score": [r2_dummy, r2_baseline_lr, r2_lr, r2_rf, r2_xgb],
```

```
    "MAE": [mae_dummy, mae_baseline_lr, mae_lr, mae_rf, mae_xgb],
    "RMSE": [rmse_dummy, rmse_baseline_lr, rmse_lr, rmse_rf, rmse_xgb]
})
# Display results
print("\n 🔹 Final Model Performance Comparison:")
print(results)
#      Save the final Data
# Save numeric_df as an Excel file
numeric_df.to_excel("Analysis Data.xlsx", index=False)
print("File saved successfully as 'Analysis Data.xlsx'")
# Download the file to computer
from google.colab import files
files.download('Analysis Data.xlsx')
```