The Future of Engineering Education
2024 Annual Conference & Exposition
Oregon Convention Center
Portland, OR . June 23 - 26, 2024
ASEE
Paper ID #44170

Causal Inference Networks: Unraveling the Complex Relationships Between Curriculum Complexity, Student Characteristics, and Performance in Higher Education

Dr. Ahmad Slim, The University of Arizona

Dr. Ahmad Slim is a PostDoc researcher at the University of Arizona, where he specializes in educational data mining and machine learning. With a Ph.D. in Computer Engineering from the University of New Mexico, he leads initiatives to develop analytics solutions that support strategic decision-making in academic and administrative domains. His work includes the creation of predictive models and data visualization tools that aim to improve student recruitment, retention, and success metrics. Dr. Slim's scholarly contributions include numerous articles on the application of data science in enhancing educational practices.

Prof. Gregory L. Heileman, The University of Arizona

Gregory (Greg) L. Heileman currently serves as the Associate Vice Provost for Academic Administration and Professor of Electrical and Computer Engineering at the University of Arizona, where he is responsible for facilitating collaboration across campus t

Melika Akbarsharifi, The University of Arizona

Melika Akbarsharifi is a Master's student in Electrical and Computer Engineering at the University of Arizona, studying under Professor Gregory L. Heileman. Her research at the Curricular Analytics Lab focuses on using machine learning and data analysis to enhance educational outcomes. Key contributions include developing a cohort-tracking analytics platform that assists in improving graduation rates by addressing curricular barriers.

Melika has co-authored papers presented at conferences such as the ASEE Annual Conference and Exposition, exploring the intersection of curriculum complexity and student performance. Her technical proficiency spans multiple programming languages and cloud computing, furthering her research into innovative educational technologies.

Kristina A Manasil, The University of Arizona

Kristi Manasil is a first-year PhD student in the School of Information at the University of Arizona. She received her bachelor's degree in Computer Science from the University of Arizona. She is interested in data visualization, machine learning, human computer interaction, learning analytics and educational data mining.

Ameer Slim, University of New Mexico

Causal Inference Networks: Unraveling the Complex Relationships Between Curricular Complexity, Student Characteristics, and Performance in Higher Education

Ahmad Slim[†], Gregory L. Heileman[†], Ameer Slim[‡],
Kristi Manasil[†], Melika Akbarsharifi[†]
{ahslim@arizona.edu, heileman@arizona.edu, ahs1993@unm.edu,
kmanasil@arizona.edu, akbarsharifi@arizona.edu}
[†]The University of Arizona
[‡]The University of New Mexico

Abstract

Numerous research studies have explored the influence of curriculum complexity on student performance, primarily focusing on factors like retention and graduation rates. Many of these investigations have employed conventional machine learning and data analysis methods, often yielding results that are challenging to interpret and convey effectively. Furthermore, these studies have generally lacked a comprehensive framework for elucidating how variables such as student gender and prior academic preparation contribute to selecting specific university programs, each characterized by its structural complexity. These studies have yet to present foundational models to elucidate the fundamental mechanisms underlying the causal relationship between the complexity of university programs, student attributes, and success metrics. Our present study introduces an innovative causal inference network model that conceptualizes the university as a dynamic system with interrelated causal relationships among its various components, encompassing students, programs, colleges, graduation rates, and more, each with their respective dependencies. This model allows us to comprehend and visually represent the direction of causality between different variables, enabling us to investigate how changes in one variable, the causal factor, impact another variable. This implementation of causality not only facilitates predictive tasks, like other conventional machine learning models (i.e., hypothetical causation), but also enables us to conduct objective "what-if" analyses (i.e., counterfactual causation) within the research context. In this study, we leverage real-world student data from 30 universities across the United States. The richness and diversity of our dataset empower us to draw robust insights into the causal relationships among various factors that influence student performance, particularly the complexity of the curriculum. A key finding from our causal analysis indicates that an increase in program complexity by 20 points is correlated with a decrease of 3. 74% in the likelihood of graduating within four years. Moreover, our counterfactual scenarios demonstrate that for students with specific demographic profiles, such as males with a certain HSGPA not receiving Pell Grants, an increase in complexity could inversely affect their graduation prospects. These nuanced discoveries underscore the importance of curriculum design in alignment with student demographics and preparation, challenging educators to balance academic rigor with the facilitation of student success. The breadth and scale of our dataset significantly enrich the quality of our conclusions, providing valuable guidance for future educational strategies and policies.

keywords: curricular complexity, causal inference, student success, graduation rates, educational data mining

1 Introduction

Curriculum complexity, an intrinsic characteristic of educational programs, has increasingly become a focal point of academic research due to its presumed impact on student performance. The architecture of a curriculum – encompassing the breadth and depth of content, the sequencing of subjects, and the interplay of various pedagogical approaches – directly influences the learning environment. This influence is often reflected in key educational outcomes such as student engagement, comprehension, retention, and graduation rates. The complexity of a curriculum, therefore, is not merely an academic concern but a pivotal factor that could shape students' educational trajectory and success. However, research in this domain often needs to be revised, particularly in unraveling the intricate causal relationships within the educational ecosystem. While numerous studies have explored how various aspects of curriculum design affect student outcomes, they predominantly employ conventional data analysis and machine learning methods. These approaches, while valuable, often lead to results that are correlative rather than causative. Therefore, the challenge lies in moving beyond identifying patterns and correlations to understanding the underlying causal mechanisms. This gap in existing research is particularly evident in the lack of a comprehensive framework that considers the multifaceted influences on student performance, including factors such as student gender, socioeconomic background, prior academic preparation, and the unique structural complexities of different university programs. Addressing this research gap, our study introduces an innovative approach – a causal inference network model – designed to conceptualize the university as a dynamic system. This model embraces the complexity of educational environments, acknowledging the myriad of interrelated components that coexist within them, such as students, academic programs, and institutional policies. Using this causal inference network model, our study aims to unravel the complex web of relationships and dependencies within the university setting. This approach is not limited to predictive capabilities, as seen in traditional machine learning models; it also enables us to engage in objective "what-if" analyses. These analyses delve into counterfactual reasoning, allowing us to explore hypothetical scenarios and their potential impacts on student outcomes. We aim to utilize this model to better understand the causal relationships between curriculum complexity and student performance metrics. By doing so, we aim to contribute a novel perspective to educational research discourse, offering theoretical insights and practical implications for curriculum design and student success strategies. This study not only seeks to fill a critical gap in current academic research but also aspires to provide educators and policymakers with a robust tool for informed decision-making, ultimately enhancing students' educational experience and outcomes.

2 Background and Literature Review

Transitioning from the Introduction, which sets the stage for the importance of understanding the influence of curriculum complexity on student performance, we dive deeper into the nuances of this relationship. This exploration is critical as it sheds light on the multifaceted nature of education and its impact on learners. When examined closely, curriculum complexity reveals itself in two primary dimensions: structural and instructional. The structural complexity entails the organization of the curriculum, focusing on aspects such as the sequencing of courses and the prerequisites required. Such structural aspects significantly influence the educational trajectory of the students. For instance, a curriculum that demands a rigid progression of courses with tightly interlinked prerequisites may inadvertently create barriers to student progression. This complexity can lead to a higher probability of student dropout or an extended time required to complete their studies. This phenomenon is well-documented in the literature, investigating how various structural designs of curricula either facilitate or hinder student advancement[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15]. On the other side of the spectrum lies instructional complexity, encompassing the teaching methodologies and support mechanisms integrated within the curriculum. This dimension concerns the depth and variety of teaching methods, the rigor of content delivered, and the support systems in place to guide students through their educational journey. Studies have indicated that while a richly diverse instructional approach can enhance learning, it may also lead to challenges if not adequately supported. The survey by Hiebert et al. highlights the potential negative impact of high instructional complexity on student engagement and comprehension[16]. When students are confronted with overly complex instructional methods with sufficient support, they can significantly improve their ability to engage and understand the material. The delicate balance between these two dimensions of curriculum complexity is essential. Achieving an optimal curriculum design involves finding a harmony where the structural and instructional complexities complement rather than contradict each other[17]. A curriculum that is overly complex in structure but lacking in instructional richness may lead to student overwhelm. In contrast, a curriculum rich in instructional complexity but needs to be better organized can create clarity and disengagement. Understanding the implications of curriculum complexity on student performance is a multifaceted endeavor. It involves considering not only the individual elements of curriculum design but also the interplay between these elements. This understanding is crucial for educators and curriculum designers as they strive to create educational experiences that are both challenging and supportive, thereby enhancing student learning outcomes and overall academic success. As we explore these dimensions of curriculum complexity, it becomes apparent that traditional methods of analysis, particularly those used in machine learning and data analytics, need to be revised to fully capture the dynamic nature of educational environments. While adept at identifying patterns and correlations, these conventional methods often need help understanding the complex causal relationships in academic settings. This leads to a critical analysis of these limitations and the need for more nuanced approaches in educational research. In educational research, conventional machine learning and data analysis methods have provided substantial insights into student performance and curriculum effectiveness. However, its utility in establishing causality within the dynamic and multifaceted world of education still needs to be improved. This limitation comes primarily from their focus on identifying correlations rather than causative relationships. Such methods are adept at sifting through large datasets to find patterns but often fail to discern whether one variable is the cause of changes in another[18]. This is particularly critical in education, where understanding the cause-and-effect relationships is vital to effective curriculum design and pedagogical strategies. Educational environments are inherently complex and characterized by a variety of interacting variables. Conventional data analysis methods, with their linear and isolated approaches, often fail to capture this complexity, leading to incomplete or even misleading insights. This limitation in effectively handling the intricate nature of educational systems is echoed in Shmueli's work, highlighting the challenges in establishing causality in statistical analysis[19]. In addition, these conventional methods often hinder the generation of actionable insights crucial for developing educational policy and curriculum. The insights from such analyses may indicate trends within educational settings but need more clarity on why these trends occur and how they can be effectively addressed. This gap in practical applicability is of great concern, as noted in the work by Daniel KB, which discusses the potential and limitations of big data and learning analytics in higher education[20]. Recognizing these limitations, there is a growing need for more comprehensive analytical approaches in educational research. Such approaches should integrate the ability to analyze complex, multidimensional data while providing a deeper understanding of the causal relationships within educational settings. This evolution in research methodologies is essential for developing more effective, evidence-based educational policies, curricula, and teaching strategies. Transitioning from the discussion of conventional machine learning and data analysis methods, it becomes clear that educational research requires more sophisticated analytical approaches, particularly in understanding the causal dynamics within educational settings. This necessity has led to the increasing importance of causal inference in academic research, representing a significant shift from mere correlation-based analysis to a focus on uncovering cause-and-effect relationships. Causal inference offers a more refined lens through which researchers can examine how different elements within the educational spectrum, such as curriculum design, instructional methods, and other environmental factors, directly impact student outcomes. This approach is essential in disentangling the complex interplay of variables within educational settings. Researchers can move beyond surface-level associations by employing causal inference methodologies to uncover the underlying mechanisms that drive educational outcomes. Adopting causal inference in educational research marks a critical advance in the field. It allows for a more comprehensive approach to analyzing education systems' multifaceted and dynamic nature. Pearl and Mackenzie's landmark work delves into the foundational concepts of causal inference, providing a framework that is increasingly being applied in educational research[21]. Their work has been instrumental in introducing causal models and the concept of counterfactuals, which are essential for understanding what might happen under different educational scenarios. Moreover, the research conducted by Holland further clarifies the distinction between observational and experimental data in drawing causal conclusions[22]. This distinction is crucial in educational settings, where randomized controlled trials are often impractical or unethical, thus necessitating sophisticated methods to draw causal inferences from observational data. Causal inference in education also aligns with the principles outlined in Morgan and Winship's work, emphasizing statistical methods for causal analysis in social sciences[23]. Applying these principles to educational research has opened new avenues for understanding how changes in specific curricula or teaching methods can affect student learning and achievement. This shift towards causal inference underscores the necessity for methodologies that accommodate the complexity of educational systems and unravel the causal relationships within them. By adopting these advanced approaches, academic researchers and policymakers can gain more nuanced insights into how different factors contribute to educational outcomes, leading to more effective and targeted interventions. In essence, the movement toward causal inference in educational research is a response to the limitations of traditional analytical methods. It offers a pathway to more robust, evidence-based insights that can profoundly impact educational practices and policies.

3 Methodology

Transitioning from the importance of causal inference in educational research, this section of this study is crucial in understanding how these advanced techniques are applied to real-world data. The dataset employed in this study comprises a rich and diverse collection of student data from 30 different universities. This data set includes several covariates or variables integral to understanding the educational landscape and student outcomes.

3.1 Data Description

The dataset features a range of variables designed to capture the multifaceted nature of student experiences and outcomes across various universities. These variables include:

1. Program_Complexity: This is a discrete variable reflecting the complexity of each program that students attend at a given university. The complexity metric could encompass factors like the number of required courses, the depth of course content, and the sequencing of course material[1,12].

2. University: A categorical variable listing the names of the universities where students are enrolled. This variable allows for analyzing how institutional differences impact student outcomes.

3. grad4: A binary variable that indicates whether a student graduated within a four-year time frame (1) or took more than four years (0). This variable is essential to assess the efficiency and effectiveness of university programs.

4. HSGPA: High School GPA for each student, measuring academic performance prior to university enrollment. This variable is often used in educational research to control for prior academic achievement, as noted in studies like[24].

5. Gender: A binary variable (0 for male and 1 for female) representing the Gender of the student. Gender has been shown to play a role in educational outcomes and choices, as explored in[25].

6. Pell_Award: A binary variable indicating whether a student received Pell Grant money (1) or not (0). This variable indicates students' socioeconomic status and is crucial to understanding access to educational opportunities and resources.

3.2 Causal Inference Network Model

The core of the study's methodology is developing and applying a causal inference network model. This model is designed to visualize and analyze the causal relationships among the various components of the university setting, including student characteristics, program features, and educational outcomes. The model enables the identification of potential causal paths and the estimation of the effects of changes in one variable on others. For instance, it can help determine how modifications in program complexity affect graduation rates or how socioeconomic factors like Pell Grant awards influence student success. The development of this model follows the principles outlined in Pearl's work, which provides a comprehensive framework for constructing and analyzing causal models in various domains, including education[26]. This model employs advanced statistical techniques to draw causal inferences from observational data, providing deeper insights than traditional correlation-based studies. By integrating these rich data and advanced modeling techniques, the study aims to uncover the nuanced interactions and causal mechanisms that influence student outcomes in higher education. This approach marks a significant contribution to the field, offering theoretical and practical insights into the factors driving student success and program effectiveness in university settings.

4 Data Analysis and Results

In the preliminary analysis phase of this educational research, the focus is on meticulously preparing the student data for comprehensive analysis. This phase encompasses several critical steps in data management, each contributing significantly to the overall integrity and accuracy of the study. Initially, the process begins with carefully reading and importing raw student data, including variables like Program_Complexity, University, grad4, HSGPA, Gender, and Pell_Award. The accuracy in this initial step is crucial as it sets the foundation for all subsequent analyses. One of the first tasks in data preprocessing is handling missing values. Missing data can lead to biases if not appropriately addressed. Depending on the data's nature and the analysis type, strategies like imputation or listwise deletion are often employed. Enders' work provides extensive methodologies for managing missing data in research[27] . Encoding categorical variables is another vital step. Variables such as University or Gender often require conversion into numerical formats for compatibility with most statistical analysis algorithms. This encoding ensures that these categorical variables are accurately interpreted in later analysis stages. The techniques and best practices for this can be found in[28]. Data normalization is also undertaken to ensure that variables measured on different scales contribute equally to the analysis. This process is essential to prevent biases in the model's performance due to variables with higher magnitudes. For more, Kuhn and Johnson's work offers insights into the importance and methods of feature scaling and normalization[29]. The culmination of these processes results in a clean, well-structured dataset that is primed for detailed analysis. This dataset forms the basis for the study's more complex analyses, underscoring the importance of thoroughness and precision in this preliminary phase.

4.1 Causal Model Building

Building on the meticulous groundwork laid by the preliminary analysis, the study advances into a pivotal phase of Causal Model Building, employing the PC (Peter-Clark) algorithm. This algorithm is instrumental in the causal discovery process, providing a means to discern potential causal structures within the dataset, a task especially pertinent in the intricate field of educational research. The PC algorithm, a linchpin in causal discovery, is adept at navigating scenarios where causal relationships are not predefined or hypothesized. It systematically tests for conditional independencies among variables, constructing a causative narrative. This aspect of the PC algorithm is particularly beneficial in educational research, where variables are often interdependent and com-

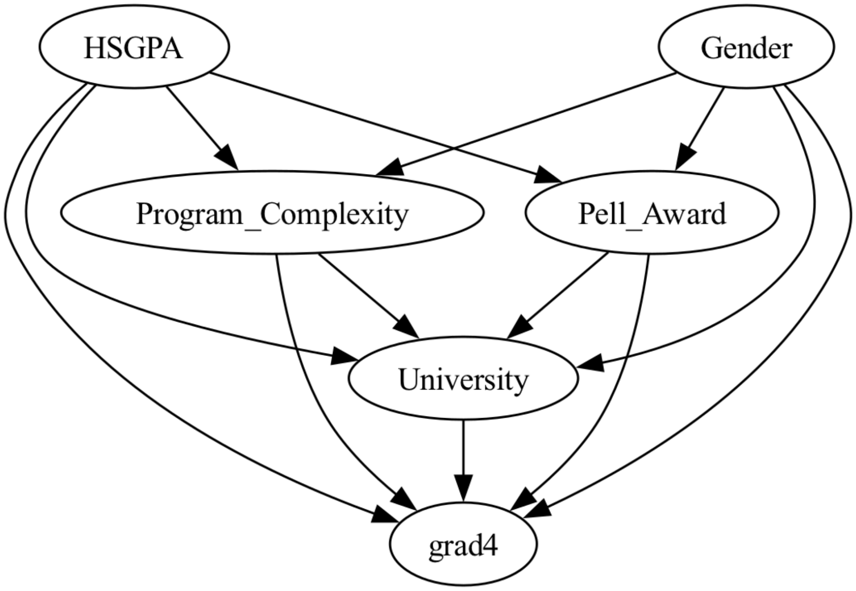

Figure 1: Causal Network Diagram

plex, as noted in[30]. When applying this algorithm in educational settings, intricate interactions such as those among Program Complexity, HSGPA, Gender, and Pell Awards are elucidated. This methodology could, for example, reveal unanticipated causal pathways, such as the impact of socioeconomic factors such as Pell grants on academic performance, going beyond direct academic indicators like HSGPA. This aligns with the foundational concepts in causal inference discussed in Pearl's work[26].

Incorporating the results of the PC algorithm into the study, we have a visual representation of the causal relationships within the educational data, as depicted in the causal network diagram shown in Figure 1. This diagram, a direct output from the PC algorithm, illustrates the potential causal pathways among key variables such as HSGPA, Gender, Pell_Award, Program_Complexity, University, and grad4. The diagram indicates, for instance, how Program_Complexity may directly or indirectly influence whether students graduate within four years (grad4) and how other variables like University or Pell_Award potentially moderate this relationship. This visual tool is instrumental in making sense of complex interactions and guiding the subsequent analysis, ensuring that the model considers the multifaceted influences on graduation rates. The causal network diagram also suggests potential mediation effects, such as how the influence of Gender on graduation rates may be mediated through variables like Pell_ Award or Program_Complexity. Integrating this causal network diagram into the study's findings provides a powerful illustration of the causal relationships suggested by the PC algorithm. It offers a concrete foundation for further analysis, such as applying statistical techniques such as logistic regression or gradient-boosting classifier (GBC) to estimate the causal effects quantitatively. These methods can validate the suggested pathways and assess the strength and significance of the relationships depicted in the diagram. By adding this visual component, the study gains a clearer understanding of the data structure and ensures that the interpretation of the causal relationships is grounded in empirical evidence. This approach aligns with the principles of causal discovery and the foundational concepts of causal inference outlined in[30,26]. Furthermore, it reinforces the importance of robust methodological frameworks for causal inference in educational research as emphasized in[23]. However, implementing the PC algorithm comes with its challenges. It requires careful navigation through the complexities of extensive and multifaceted educational datasets, demanding computational rigor and a profound understanding of the underlying educational phenomena. The algorithm's adaptability for exploratory research, where relationships aren't preconceived, is a significant advantage. Despite its strengths, the PC algorithm has limitations, notably in situations with hidden confounders or within large, high-dimensional datasets. These scenarios require a nuanced interpretation of the resulting causal models, considering the impact of data quality, including sample size and measurement accuracy, on the findings. Integrating the PC algorithm into this study is part of a broader initiative to infuse more rigorous, data-driven causal analysis into educational research. This method aligns with the escalating focus on evidence-based educational policy and practice, aiming to uncover complex causal structures within educational data, as expounded in[23]. This comprehensive approach contributes to a more nuanced understanding of educational processes and outcomes, informing precise and effective interventions. In essence, deploying the PC algorithm for building a causal model embodies an advanced and refined approach in educational research. It holds the potential to unlock novel insights into the causal dynamics that shape educational environments and student outcomes, marking a significant stride in the field.

4.2 Exploratory Data Analysis

Following the intricate causal model constructed using the PC algorithm, our study's exploratory data analysis (EDA) phase probes deeper into the causal network diagram, elucidating the multifaceted relationships it represents. This phase is instrumental in visualizing the distribution of program complexity across various educational institutions, evaluating gender disparities within these institutions, and examining the association between program complexity and graduation rates. A Kernel Density Estimate (KDE) analysis, as depicted in Figure 2, reveals a discernible pattern in which students with higher High School GPA (HSGPA) scores tend to gravitate towards more complex academic programs. This pattern suggests a correlation in which students' educational background influences their choice of academic challenges[31]. The KDE analysis compares the program complexity preferences of two cohorts of male students delineated by their HSGPA. The first cohort, with HSGPAs ranging from 2.0 to 2.5, presents a different complexity preference than the second cohort, which consists of students with HSGPAs greater than 3.5. These KDE plots underscore the varying academic dispositions based on prior performance, indicating that students with higher academic achievements in high school are more inclined to enroll in structurally and academically more demanding programs. This observation could imply that students with higher HSGPA are more confident in their academic capabilities or are better prepared to handle the rigors of complex programs. Such insights are crucial, as they point toward underlying educational stratification mechanisms where academic proficiency influences program selection, potentially leading to resource concentration in certain student demographics[32]. Furthermore, this propensity of students with higher academic achievement to opt for more challenging programs may reflect the selective nature of such programs, which may prioritize applicants with a solid educational foundation. Thus, comparative KDE analysis provides a statistical lens through which we can gauge the probability density of students enrolling at varying levels of program complexity, laying the groundwork for understanding how educational policies and interventions could be tailored to bridge the academic preparedness gap and foster equitable access to challenging curricula[33].

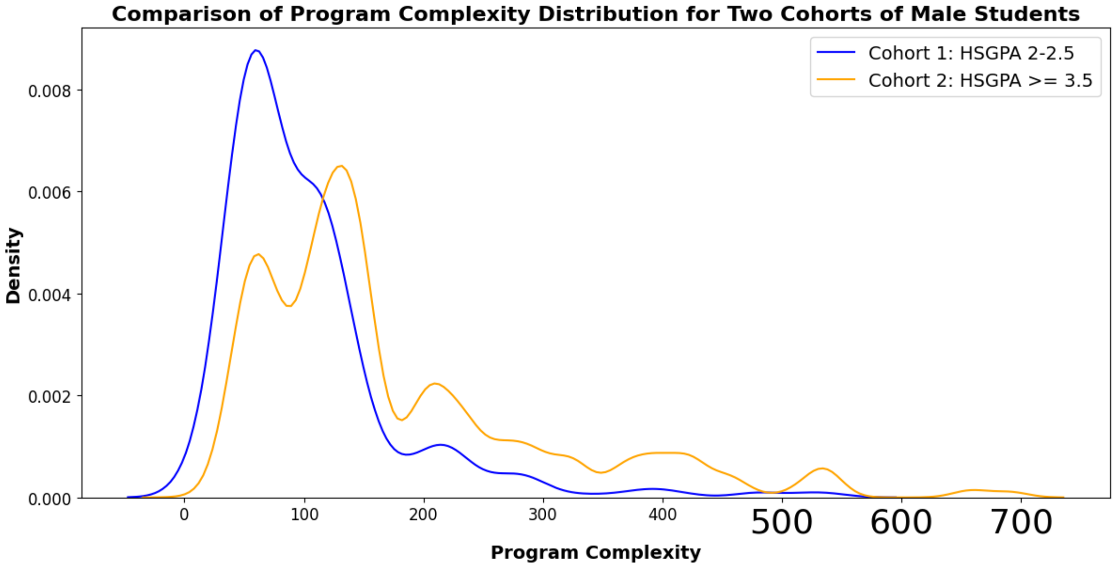

Figure 2: Comparative KDE of Program Complexity for male students with different HSGPA ranges.

Continuing from the insights provided by the KDE analysis, we further examine the variability in program complexity among universities. This part of the exploratory data analysis focuses on how the structural aspects of university curricula influence student enrollment decisions. As highlighted in Figure 3, the distribution of program complexity varies notably between different institutions, such as University '1' and University '3'. This variability is not merely incidental but indicative of these institutions' diverse academic cultures and curricular frameworks. The KDE plot for University '1', with a multi-peaked distribution, suggests a curriculum that offers a wide array of programs ranging from less to more complex. In contrast, University '3's distribution, characterized by fewer peaks, might imply a more focused or streamlined set of programs. The complexity of a program often reflects the depth and breadth of the curriculum, the number and sequencing of courses, and the level of integration of cross-disciplinary studies, all of which contribute to the academic experience a university offers[34]. The differences in program complexity distributions can also be tied to each university's unique mission and educational philosophy. Some institutions may prioritize a liberal arts education, offering a broad range of courses with varying levels of complexity. In contrast, others may focus on specialized or professional programs that are inherently more complex due to their depth in a particular field[35]. Moreover, the choice of a specific university program is not only influenced by the interests and academic capabilities of students but also by the perceived value of the program in the job market, which can be linked to its complexity[36]. Students might perceive more complex programs as providing a competitive edge for future employment, thus influencing their enrollment decisions. This variability in program complexity among universities has significant implications for educational policy and student guidance. Universities need to ensure that the complexity of their programs is aligned with their educational objectives and the needs of their student population. This alignment is critical for maintaining academic standards while promoting student success and retention[37]. The insights from this analysis underscore the need for a nuanced understanding of how program offerings relate to student choice and success. They provide a foundation for further research into optimizing program complexity to enhance educational outcomes and student satisfaction.

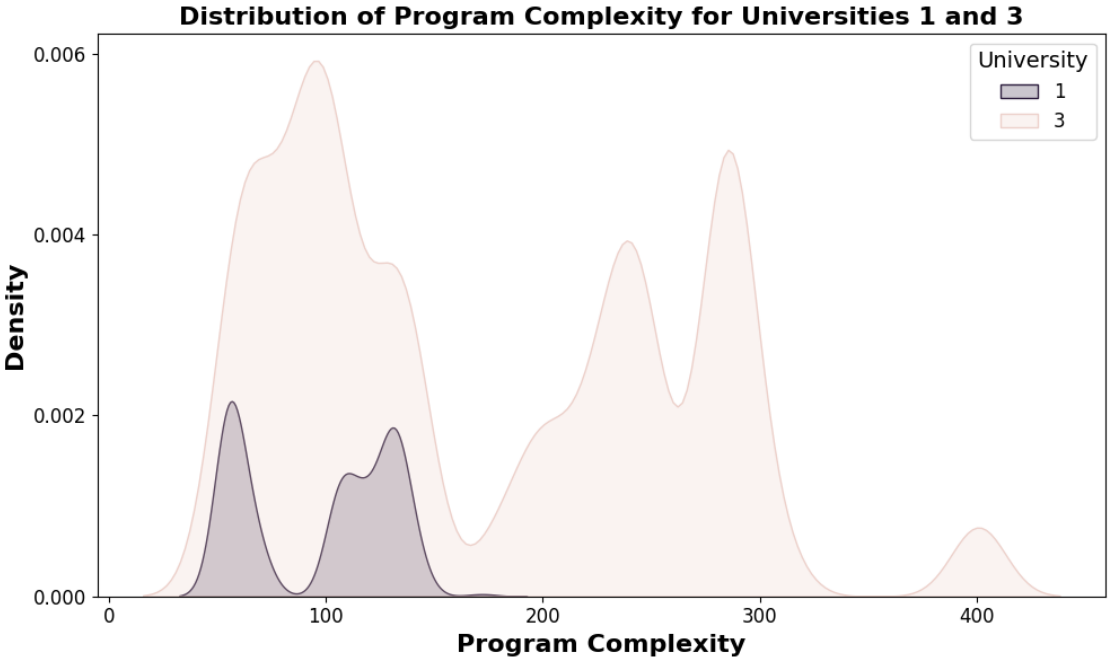

Figure 3: Distribution of Program Complexity for Universities 1 and 3.

The exploratory data analysis extends to the domain of financial aid, specifically examining the distribution of Pell Grant awards in relation to Gender. As evidenced in Figure 4, there is a marked gender disparity in the allocation of Pell Grants, with a higher proportion of female students being recipients. This finding is consistent with national trends that show women are more likely to receive Pell Grants, a phenomenon that has been the subject of substantial research in higher education policy[38]. The Pell Grant program, designed to provide need-based grants to low-income undergraduate students, is crucial in facilitating access to higher education. The gender disparity observed may reflect underlying socioeconomic factors that affect men and women differently. Research suggests that women are more likely to come from backgrounds that qualify for financial aid, and they also tend to apply for it at higher rates than their male counterparts[39]. The overrepresentation of female students among Pell Grant recipients could also be linked to broader educational attainment patterns. Women have been enrolling in and graduating from higher education institutions at higher rates than men, which may correlate with a greater need for financial support to complete their education[40]. Furthermore, the trend may be amplified by differences in the field of study choices, with women often entering less lucrative fields immediately after graduation, necessitating financial assistance during their studies[41]. This gender-based analysis of financial aid not only highlights the disparities that exist in higher education financing but also raises important policy considerations. For instance, there may be a need to examine the support structures for male students, particularly those from low-income backgrounds, to ensure equitable access to financial aid resources. Understanding these gender disparities is critical for higher education institutions and policymakers as they strive to design financial aid policies that effectively address the needs of all students. Such policies can have far-reaching implications for promoting diversity and inclusion within the educational system.

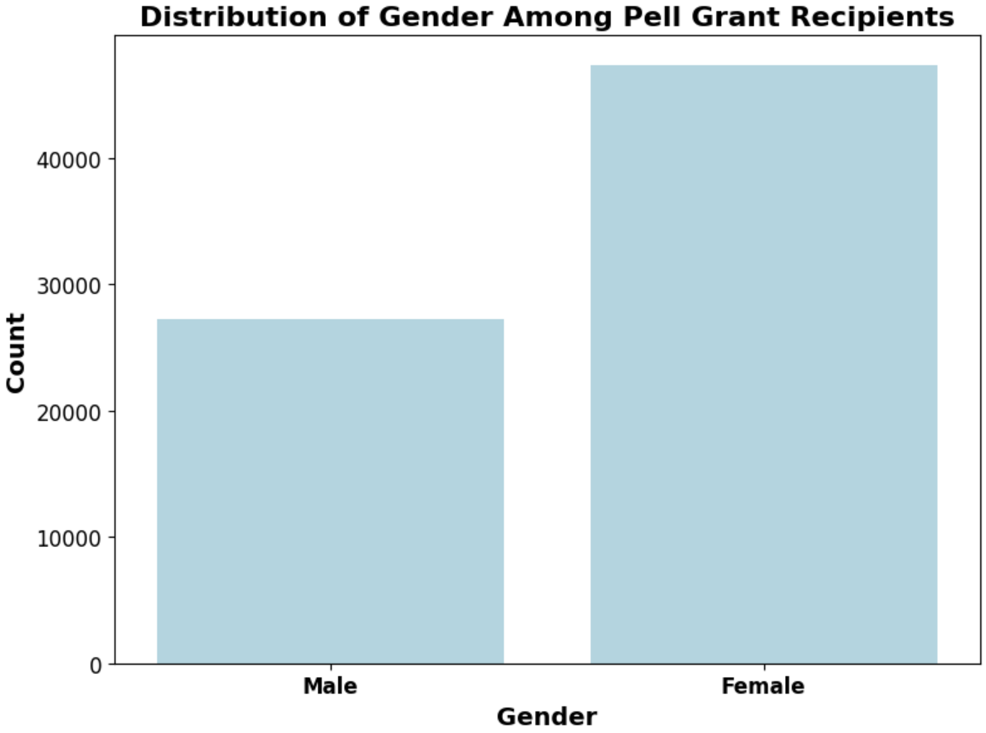

Figure 4: Gender distribution among Pell Grant recipients.

The distribution of Pell Grants, a proxy for the socioeconomic status of the student body, varies significantly across different universities. In Figure 5, this variation is starkly evident in the comparison between University ’3’ and University ’6’, with the former showing a more equitable distribution and the latter exhibiting a higher concentration of Pell Grant recipients. This discrepancy may reflect the diverse economic backgrounds of the students they serve, suggesting that University ’6’ may cater to a more economically disadvantaged population. The Pell Grant program is a federal initiative in the United States that provides need-based grants to low-income students to promote access to postsecondary education. The distribution of these grants can be influenced by several factors, including the institution’s mission, the demographic it attracts, and the geographic region it serves. For example, some universities might be situated in areas with higher levels of economic hardship or have missions that focus on serving underprivileged communities, which could explain a higher concentration of Pell Grant recipients[42]. Furthermore, the contrasting Pell Grant distributions could also be an outcome of the differing institutional aid policies and the capacity of these universities to provide financial assistance beyond federal aid. Universities with robust financial aid programs might be able to supplement the needs of low-income students more effectively, thus attracting a larger population eligible for Pell Grants[43]. Additionally, the observed discrepancies might relate to the academic programs offered by the universities. Some academic fields may draw a demographic that is more likely to qualify for Pell Grants, reflecting the intersection between students’ socioeconomic background and their field of study choice[44]. The insights gleaned from the analysis of Pell Grant distributions are significant for understanding how universities can develop targeted strategies to support their student populations. They also highlight the importance of considering the interplay between federal, state, and institutional financial aid policies and their collective impact on promoting access and equity in higher education. Understanding these institutional differences is crucial for policymakers who are tasked with ensuring that higher education remains accessible to all students, regardless of their economic background. It also serves as a call to action for universities to examine their policies and practices to ensure they meet the needs of their diverse student bodies.

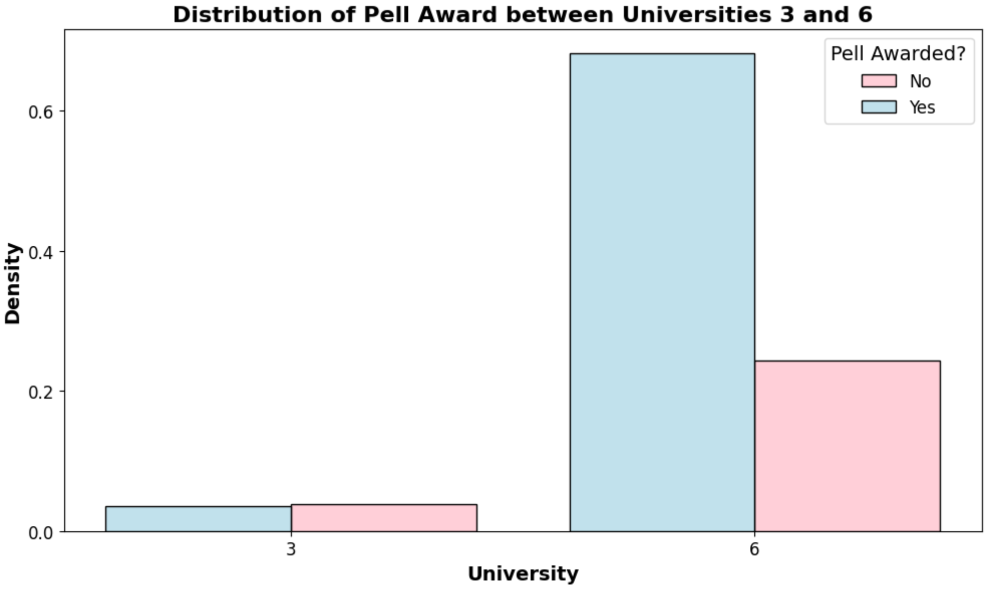

Figure 5: Comparative Distribution of Pell Grants between Universities 3 and 6.

The variability in educational outcomes is further accentuated when we scrutinize the 4-year graduation rates across different universities. As depicted in Figure 6, there is a discernible difference in the graduation efficiency between University '5' and University '7'. University 5' has a commendably higher proportion of students graduating within a standard four-year period than University 7', which suggests disparities in educational efficiency and perhaps differences in the support structures provided to students. The 4-year graduation rate is a critical metric for assessing the effectiveness of universities in facilitating timely degree completion. Factors influencing this rate include the rigor and complexity of academic programs, the level of educational support, and the financial resources available to students[45]. For instance, University '5's higher graduation rate may be attributed to more integrated academic advising systems, compelling first-year experiences, or greater availability of financial aid that minimizes the need for students to work while studying, allowing them to focus on their studies and complete their degrees on time[40]. Conversely, University '7's lower 4-year graduation rate might reflect challenges such as inadequate academic advising, insufficient financial aid, or a student body with more external commitments, which could extend the time needed to graduate[46]. It's also possible that University '7' serves a larger proportion of nontraditional students, who often take longer to graduate due to part-time enrollment, work, and family responsibilities[47]. These findings highlight the importance of institutional accountability in promoting student success. Universities are encouraged to examine their policies and programs to identify barriers to timely graduation and develop interventions to improve educational efficiency[48]. Understanding these graduation rates in the context of institutional characteristics provides valuable insights for educational leaders and policymakers aiming to enhance student success and institutional performance. It also underscores the need for nuanced data analysis beyond simple comparisons to understand the factors contributing to these outcomes.

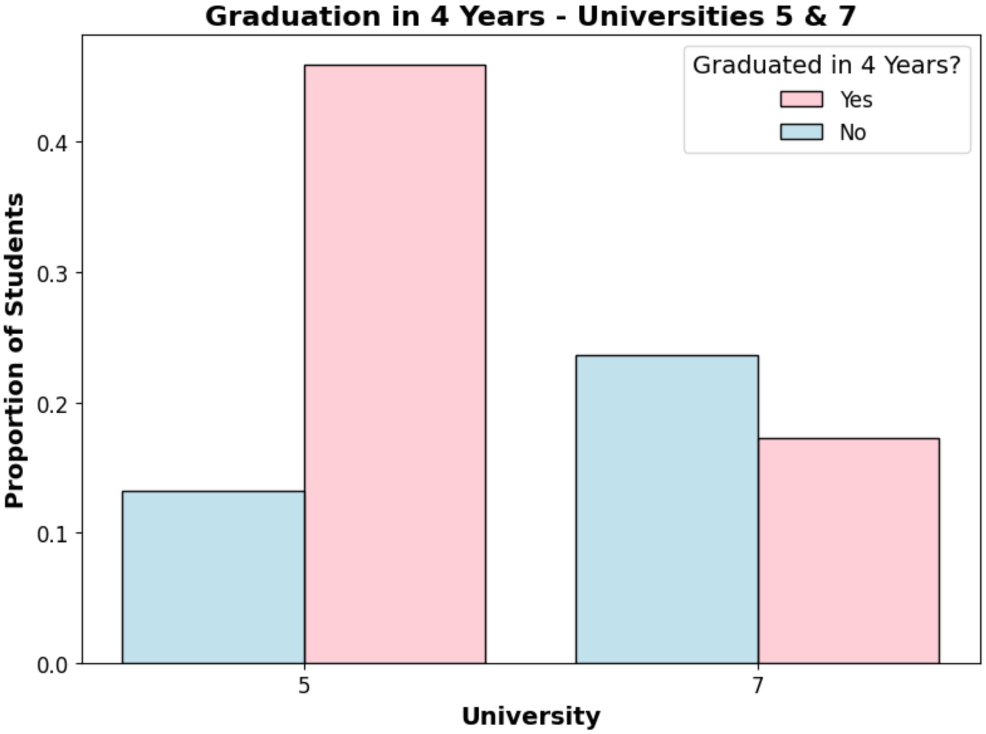

Figure 6: 4-Year Graduation Rates for Universities 5 and 7.

Exploring the influence of program complexity on the 4-year graduation rate reveals a nuanced aspect of academic dynamics. Figure 7 illustrates a salient feature of higher education—the existence of a complexity threshold that, when surpassed, corresponds with a decreased likelihood of graduating within the traditional four-year window. This phenomenon is discernible across student cohorts, each exhibiting a critical complexity point beyond which graduation rates falter. Program complexity, as measured by the breadth and depth of coursework, the integration of interdisciplinary studies, and the academic load, can pose significant challenges to students. The analysis suggests that there is an optimal balance to be struck where program rigor enhances learning without impeding progress toward graduation. This aligns with the body of research that underscores the delicate interplay between academic challenge and student support systems[49]. For both cohorts analyzed, a program complexity threshold suggests that institutions must carefully consider the design of their curricula. Programs must be rigorous enough to provide a quality education yet structured to promote timely completion. This underscores the need for universities to continually assess and recalibrate their curriculum to align with student capabilities and institutional support services[50]. Furthermore, this relationship between program complexity and graduation rates calls attention to the importance of academic advising and support services. Adequate advising can help students navigate complex programs effectively, making informed decisions about course loads and sequences that align with their circumstances and academic goals[51]. The findings from this study contribute to the discourse on curriculum development and student success, emphasizing that higher education institutions should focus not only on the content of education but also on the structure and delivery of their programs. By doing so, they can create pathways to degree completion that accommodate a diverse student body, each with unique academic and personal backgrounds.

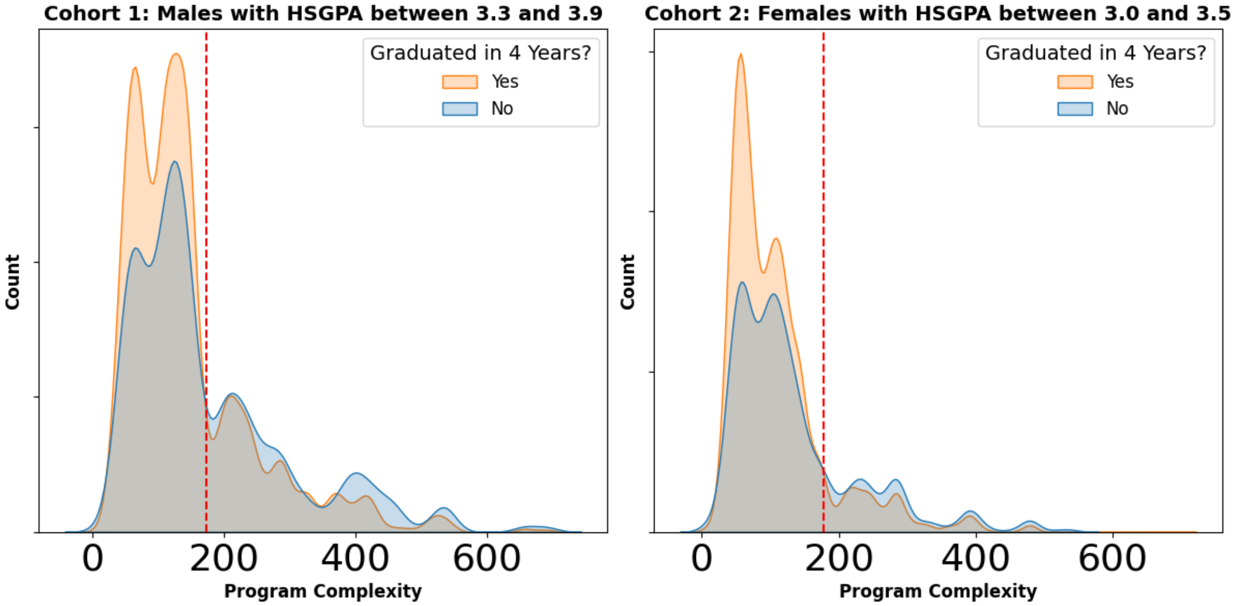

Figure 7: 4-Year Graduation Rates by Program Complexity for two student cohorts.

The exploratory data analysis conducted across various dimensions of the educational experience synthesizes a multifaceted narrative about the determinants of student success. The juxtaposition of program complexity with graduation rates, the distribution of financial aid across gender lines, and the variability in institutional support all contribute to a complex mosaic of factors affecting educational outcomes. These insights, gleaned from the visualizations provided by Figures 2 through 7, are not merely descriptive but form the basis for a deeper inquiry into the mechanics of student success. The KDE plots and bar charts are more than illustrative tools; they visually represent the underlying data that capture trends, outliers, correlations, and anomalies. The narrative that emerges from these figures points to the interdependence of student demographics, such as Gender and socioeconomic status, with institutional characteristics like program offerings and support structures. For instance, the overrepresentation of female students among Pell Grant recipients could reflect broader societal trends, such as the gender gap in economic status and the drive towards higher education as a pathway to economic mobility[52]. Similarly, the comparison of graduation rates between institutions, as shown in Figure 6, underscores the need for policies that recognize and address the diversity of student experiences. Universities like University ’5’, which exhibit higher 4-year graduation rates, may provide models of institutional practices that effectively support timely degree completion. Understanding what these institutions do differently can inform broader policy initiatives to improve graduation rates across the board[53]. Furthermore, the critical program complexity thresholds identified in Figure 7 highlight the importance of curriculum design in student retention and success. They suggest a need for curricular pathways that are challenging yet navigable, with academic advising and support tailored to help students manage their course loads without compromising their time-to-degree[54]. These findings contribute significantly to our understanding of the educational landscape and have important implications for future research and policy-making. They suggest that efforts to improve educational outcomes should be holistic, considering the academic content and the context in which education is delivered. As such, they provide a compelling argument for integrated policy responses that address the economic, educational, and institutional barriers to student success[55].

4.3 Causal Effect Estimation

In this pivotal section of our analysis, logistic regression and GBC are employed to estimate the causal effect of curriculum complexity on 4-year graduation rates, incorporating key confounders such as High School GPA (HSGPA) and Gender. These methodologies enable us to elucidate the complex causal pathways influencing educational outcomes. Logistic regression, a widely utilized statistical method for binary classification problems, is applied to model the likelihood of students graduating within four years based on curriculum complexity while controlling for HSGPA and Gender. This approach is particularly effective in isolating the effect of curriculum complexity on graduation rates, accounting for the potential influence of these confounders[56]. The GBC, a powerful ensemble machine learning technique, further refines our analysis. Known for its high predictive accuracy, this method aggregates weak predictive models to form a robust predictor, making it highly suitable for our complex educational dataset[57]. To estimate the causal impact, we adopt a counterfactual framework, expressed mathematically as

$$E_{H,G}\left[E(Y \mid do(t=1), H, G) - E(Y \mid do(t=0), H, G)\right] \tag{1}$$

This equation calculates the expected value, over confounders HSGPA (H) and Gender (G), of the difference in the conditional expected value of the graduation outcome (Y) given the do-operator interventions $do(t = 1)$ and $do(t = 0)$ where t represents the treatment variable (curriculum complexity, in our case). Essentially, this formula allows us to evaluate the expected change in graduation rates if a student were subjected to varying levels of program complexity, controlling for their High School GPA and Gender[58]. By applying these sophisticated statistical and machine learning tools, our study moves beyond correlation to unravel the causal relationships that govern education. These techniques provide invaluable insights into how various factors, including program complexity, High School GPA, and Gender, collectively shape students' likelihood of graduating in four years, offering crucial educational policy and curriculum design guidance.

4.3.1 Interpretation of Results

In the context of our study's causal effect estimation, the results obtained from logistic regression and GBC provide a nuanced view of the impact of curriculum complexity on 4-year graduation rates. The logistic regression model yields a causal estimate of -0.000469, while the GBC model provides an estimate of -0.000623. These negative values indicate that as program complexity increases, the probability of students graduating within four years slightly decreases, holding other variables constant. The interpretation of these findings is nuanced. The negative coefficient from the logistic regression analysis suggests a small, yet statistically significant, decrease in the likelihood of on-time graduation as program complexity rises. This could imply that overly intricate programs may hinder some students' ability to graduate in a standard timeframe, potentially due to the increased academic demands that come with complexity. Similarly, the GBC model's estimate reinforces this notion but indicates a slightly more significant effect. Given the non-linear capabilities of the GBC model, this might reflect a more nuanced relationship between program complexity and graduation rates, capturing impacts that the logistic regression may not fully detect due to its linear nature. These results, while subtle in magnitude, are crucial for higher education decision-makers. They imply that curriculum complexity has a tangible, if modest, impact on graduation timelines. University administrators and policymakers should consider these findings when designing or restructuring academic programs. The goal would be to balance maintaining academic rigor and ensuring that complexity does not become a barrier to graduation. For instance, curricula might be designed with tiered levels of complexity, allowing students to choose pathways that align with their academic preparation and life circumstances. Similarly, advising services could be tailored to help students navigate complex requirements, and support programs could be implemented to assist students who may struggle with more demanding courses. Given the stakes involved in graduation rates—for students' futures, institutional accountability, and the broader workforce—the insights from this causal analysis are non-trivial. They can inform targeted interventions aimed at smoothing the path to graduation, which could profoundly impact individual and institutional success.

The culmination of this study is the integration of rigorous data preprocessing, sophisticated causal model building, exploratory data analysis, and the application of advanced statistical techniques. Together, these methodologies form a comprehensive approach to elucidate the causal relationships present within educational data. By meticulously preparing the dataset and employing robust statistical models, we have isolated and quantified the effects of curriculum complexity on student outcomes, particularly 4-year graduation rates. The causal model, developed through the PC algorithm, laid the groundwork for understanding the intricate relationships between various educational variables. Following this, exploratory data analysis provided a visual and statistical exploration of key factors such as program complexity, gender disparities, and the distribution of financial aid. Advanced statistical techniques, including logistic regression and GBC, were then applied to estimate the causal impact of curriculum complexity on graduation rates, revealing that increased complexity slightly diminishes the likelihood of graduating within four years. These findings, as evidenced by the negative causal estimates obtained, underscore the delicate balance that higher education institutions must maintain between offering rigorous academic programs and ensuring students can complete these programs promptly. The study's insights are particularly relevant for university administrators and policymakers tasked with designing curricula and support systems that foster student success. The implications extend beyond academia, as graduation rates directly impact the workforce and society at large. Institutions may need to re-evaluate their curriculum structures and consider the introduction of support mechanisms that assist students in navigating complex programs[59,60]. Furthermore, these insights could guide future research, informing longitudinal studies that track the long-term impact of curricular changes on student outcomes. As Long[61] indicates, addressing academic barriers in higher education is essential for broadening access and improving graduation rates. The nuanced understanding of the factors affecting graduation rates can also inform the development of online learning platforms, which have been shown to influence student course outcomes significantly[62]. The study provides a data-driven foundation for initiatives to improve educational attainment, emphasizing the need for evidence-based decision-making in higher education.

4.4 Intervention and Counterfactual Analysis

In this section of our study, we advance beyond identifying causal relationships to actively manipulate these relationships within a Graphical Causal Model (GCM) framework. This model serves as a visual and mathematical representation of the causal relationships and is instrumental for conducting both intervention analysis and counterfactual reasoning. The GCM was constructed following the guidelines by Pearl, using the relationships uncovered in the PC algorithm phase[58]. This model lets us articulate the assumed causal pathways visually and formalize the relationships using structural equations mathematically. These equations form the basis for our intervention analysis, where we simulate the effect of changes in program complexity on graduation rates using the Average Treatment Effect (ATE) within the potential outcomes framework:

$$\text{ATE} = E[Y(1) - Y(0)] \tag{2}$$

where $Y(1)$ and $Y(0)$ denote the potential outcomes on graduation under treatment and control conditions, respectively. The intervention analysis involves modifying the curriculum complexity variable within our model to observe the hypothetical changes in graduation outcomes. This analysis type is often called a 'do-operation' in causal inference literature[21]. By setting the curriculum complexity variable to specific values, we can estimate the expected change in graduation rates as if the complexity had been set to those values in the real world. Through a simulated intervention that increased program complexity by 20 points and using the GBC model, our study observed a 3.74$\%$ decrease in the four-year graduation rate. This substantial reduction underscores the significant impact of curriculum design choices on student outcomes. In addition to intervention analysis, we explore counterfactual scenarios, which allow us to answer 'what-if' questions. These scenarios involve hypothesizing different outcomes based on alterations to the observed data. For instance, we might ask: "What would the graduation rate for a student with a given HSGPA and Gender have been if they had enrolled in a program with lower complexity?" Counterfactual analysis provides a deeper understanding of individual-level causation, which can be particularly useful for personalized academic advising and policy planning[23]. For example, consider a male student with a high school GPA (HSGPA) of 3.08, not receiving a Pell Grant, with a program complexity score of 154, attending University '2'. Our counterfactual analysis suggests that increasing program complexity by 20 points could change this student's four-year graduation status from successful to unsuccessful. This result is a stark illustration of the sensitivity of graduation outcomes to even moderate changes in curriculum complexity and is expressed as:

$$Y_{\text{counterfactual}}(u) = Y(u) \mid do(\text{Program\_Complexity} = \text{Program\_Complexity} + 20) \tag{3}$$

These findings have profound implications for university administrators and policymakers. The intervention analysis indicates that curriculum complexity should be calibrated carefully to ensure it does not hinder student success. The counterfactual scenarios highlight the importance of personalized academic advising; students on the margin of these complexity thresholds could be at risk of not graduating on time, and tailored support could be critical for their retention and success.

5 Discussion

This section critically examines the interplay between program complexity, student demographics, and four-year graduation rates, integrating the diverse strands of evidence presented in our analysis. Our investigation reveals that students with HSGPA tend to enroll in more complex academic programs. This finding aligns with theories of academic self-selection where students seek environments that match their preparation levels and aspirations[50]. The KDE analysis showed that students with higher HSGPAs are more likely to pursue more complex programs. This might reflect confidence in their academic capabilities or an ambition to challenge themselves, a narrative supported by the positive association in the logistic regression analysis. However, the GBC model nuanced this picture by indicating a slight negative causal effect of increased program complexity on graduation rates. This subtle yet significant finding suggests that while students with high academic achievements may seek out and initially succeed in more complex programs, there is a point where increased complexity may start to hinder their ability to graduate within the traditional four-year timeline[45]. The causal estimates from logistic regression and GBC models underscore the sensitivity of student outcomes to program complexity. Specifically, our intervention analysis showed that adding 20 points to the program complexity score could result in a 3.74% decrease in the likelihood of graduating on time, which is substantial when considering the scale of a university student body. Furthermore, the counterfactual scenarios brought to light the individual-level effects of curriculum complexity, illustrating how a seemingly moderate increase in complexity can decide a student's ability to graduate on time, especially for students who may already be near the threshold of their academic capabilities. This nuanced understanding of the relationship between curriculum complexity and graduation rates holds significant implications for educational policy and curriculum development. While pursuing rigorous academic standards is a laudable goal, our findings caution against an uncritical elevation of complexity. Instead, they suggest that universities adopt a more individualized approach to curriculum design that considers the diverse capabilities and needs of the student population[40]. The counterfactual analysis, in particular, has powerful implications for personalized academic advising and support. By understanding the specific factors affecting individual students, advisors, and educators can tailor their support to help each student navigate their academic journey more effectively[23]. Our study contributes to the ongoing discourse on how best to structure university programs to support student success. It highlights the need to balance challenging students academically and providing them with a clear path to graduation, affirming the critical role of informed, data-driven decision-making in higher education.

Our approach to understanding these relationships through a GCM and applying causal inference techniques, such as the do-calculus for intervention analysis and counterfactual scenarios, is unique in educational research. While previous studies have focused on the correlation between student characteristics and educational outcomes, our study extends this by examining the causal impact of curriculum complexity on graduation rates. This has allowed us to confirm patterns noted in prior research and understand the directional influence of program complexity on student success. Previous studies have demonstrated the importance of academic preparation and curriculum structure on student outcomes[50,40]. However, our study contributes to the field by applying a causal inference framework to quantify the effect of curriculum complexity, providing a more detailed understanding of how it influences graduation rates. This application aligns with recent methodological advancements emphasized by researchers like Pearl and Morgan and Winship, who advocate for a more rigorous causal analysis in social research[58,23]. Furthermore, our findings challenge some traditional assumptions about program complexity. Whereas some pedagogical theories posit that higher complexity might foster deeper learning and better prepare students for post-graduation challenges, our counterfactual analysis suggests a threshold beyond which additional complexity may hinder timely graduation. This nuance adds a layer of complexity to the dialogue on curriculum design, as identified by Bound, Lovenheim, and Turner, who explore the multifaceted influences on college completion rates[45]. By contrasting our findings with the broader literature, we highlight the contribution of our work to the ongoing discussion about educational attainment. Our study underscores the need for nuanced curriculum design that considers student demographics and academic backgrounds to support diverse educational needs and promote equity in higher education outcomes.

6 Conclusion

In conclusion, our study has made several key contributions to educational research by applying a rigorous causal inference framework to examine the effects of curriculum complexity on 4-year graduation rates. Using a Graphical Causal Model (GCM), we have demonstrated that program complexity has a quantifiable impact on student graduation timelines. Our intervention and counterfactual analyses indicate that an increase in complexity by 20 points is associated with a 3.74% decrease in the likelihood of graduating within four years, highlighting the delicate balance required in curriculum design. Nuanced understanding of the relationship between curriculum complexity and student success contributes significantly to this study. We have shown that while students with higher academic achievements tend to enroll in more complex programs, there is a threshold beyond which additional complexity may impede timely graduation. This finding challenges the assumption that increased complexity leads to better preparation for post-graduation challenges. Our study extends the existing literature, which primarily focuses on correlational analyses, by providing a deeper causal analysis of how curriculum structures affect educational outcomes. In doing so, we contribute to the discourse on how universities can optimize curriculum design to accommodate the diverse needs of their student body, thereby promoting equitable educational outcomes. For future research, there are several potential avenues to explore. One area would be to conduct longitudinal studies to track the long-term effects of curriculum complexity on a broader range of student outcomes, including post-graduation employment and earnings. Another area could be to investigate the impact of curriculum complexity within different fields of study or types of institutions to understand whether and how these effects vary across different educational contexts. In addition, further research could explore the effectiveness of various academic support interventions in helping students navigate complex curricula. This could include studies on the role of academic advising, tutoring, and other forms of student support services. Finally, there is scope for international comparative studies to understand how curriculum complexity impacts graduation rates in different educational systems and cultural contexts. The findings from our study underscore the importance of data-driven decision-making in higher education and the potential for causal inference methodologies to inform policy and practice in this field.

References

[1] A. Slim, J. Kozlick, G. L. Heileman, J. Wigdahl, and C. T. Abdallah, "Network analysis of university courses," in *Proceedings of the 6th Annual Workshop on Simplifying Complex Networks for Practitioners*. Seoul, Korea: ACM, 2014.

[2] A. Slim, J. Kozlick, G. L. Heileman, and C. T. Abdallah, "The complexity of university curricula according to course cruciality," in *Proceedings of the 8th International Conference on Complex, Intelligent, and Software Intensive Systems*. Birmingham City University, Birmingham, UK: IEEE, 2014.

[3] J. Wigdahl, G. L. Heileman, A. Slim, and C. T. Abdallah, "Curricular efficiency: What role does it play in student success?" in *Proceedings of the the 121st ASEE Annual Conference and Exposition*. Indianapolis, Indiana, USA: IEEE, 2014.

[4] A. H. Slim, G. L. Heileman, J. Kozlick, and C. T. Abdallah, "Employing markov networks on curriculum graphs to predict student performance," in *Proceedings of the 13th International Conference on Machine Learning and Applications*. Detroit, MI: IEEE, 2014.

[5] A. Slim, G. L. Heileman, J. Kozlick, and C. T. Abdallah, "Predicting student success based on prior performance," in *Proceedings of the 5th IEEE Symposium on Computational Intelligence and Data Mining*. Orlando, FL: IEEE, 2014.

[6] A. Slim, G. L. Heileman, E. Lopez, H. A. Yusuf, and C. T. Abdallah, "Crucial based curriculum balancing: A new model for curriculum balancing," in *2015 10th International Conference on Computer Science Education (ICCSE)*, July 2015, pp. 243–248.

[7] A. Slim, G. L. Heileman, W. Al-Doroubi, and C. T. Abdallah, "The impact of course enrollment sequences on student success," in *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, March 2016, pp. 59–65.

[8] A. Slim, G. L. Heileman, M. Hickman, and C. T. Abdallah, "A geometric distributed probabilistic model to predict graduation rates," in *2017 IEEE Cloud Big Data Computing (CBDCom)*, Aug 2017, pp. 1–8.

[9] A. Slim, D. Hush, T. Ojha, , C. T. Abdallah, G. L. Heileman, and G. El-Howayek, "An automated framework to recommend a suitable academic program, course and instructor," in *Proceedings of the Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*. San Francisco, CA, USA: IEEE, 2019.

[10] G. L. Heileman, C. T. Abdallah, A. Slim, and M. Hickman, "Curricular analytics: A framework for quantifying the impact of curricular reforms and pedagogical innovations," *CoRR*, vol. abs/1811.09676, 2018.

[11] G. L. Heileman, M. Hickman, A. Slim, and C. T. Abdallah, "Characterizing the complexity of curricular patterns in engineering programs," in *2017 ASEE Annual Conference and Exposition*. Columbus, Ohio: ASEE Conferences, 2017.

[12] A. Slim, "Curricular analytics in higher education," Ph.D. dissertation, The University of New Mexico, 2016.

[13] M. Hickman, "Development of a Curriculum Analysis and Simulation Library with Applications in Curricular Analytics," Master's thesis, The University of New Mexico, 2017.

[14] B. F. Mon, A. Wasfi, M. Hayajneh, and A. Slim, "A study on role of artificial intelligence in education," in *2023 International Conference on Computing, Electronics Communications Engineering (iCCECE)*, 2023, pp. 133–138.

[15] A. Wasfi, B. F. Mon, M. Hayajneh, A. Slim, and N. A. Ali, "Optimizing assessment placement and curriculum structure through graph-theoretic analysis," in *2023 15th International Conference on Innovations in Information Technology (IIT)*, 2023, pp. 93–97.

[16] J. Hiebert and D. A. Grouws, "The effects of classroom mathematics teaching on students' learning," in *Second Handbook of Research on Mathematics Teaching and Learning*, F. Lester, Ed. Charlotte, NC: Information Age, 2007, pp. 371–404.

[17] R. J. Marzano and J. S. Kendall, *The new taxonomy of educational objectives*. Thousand Oaks, CA: Corwin Press, 2007.

[18] V. Mayer-Schnberger, *Big Data: A Revolution That Will Transform How We Live, Work and Think. Viktor Mayer-Schnberger and Kenneth Cukier*. London, GBR: John Murray Publishers, 2013.

[19] G. Shmueli *et al.*, "To explain or to predict?" *Statistical Science*, vol. 25, no. 3, pp. 289–310, 2010.

[20] B. K. Daniel, Ed., *Big Data and Learning Analytics in Higher Education: Current Theory and Practice*, 1st ed. Springer International Publishing, 2017.

[21] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*, 1st ed. USA: Basic Books, Inc., 2018.

[22] P. W. Holland, "Statistics and causal inference," *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.

[23] S. L. Morgan and C. Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd ed., ser. Analytical Methods for Social Research. Cambridge University Press, 2014.

[24] C. R. Belfield and P. M. Crosta, "Predicting success in college: The importance of placement tests and high school transcripts," Community College Research Center, Teachers College, Columbia University, Flushing, NY, CCRC Working Paper 42, Feb 2012.

[25] C. Buchmann, T. A. DiPrete, and A. McDaniel, "Gender inequalities in education," *Annual Review of Sociology*, vol. 34, no. 1, pp. 319–337, 2008.

[26] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. USA: Cambridge University Press, 2009.

[27] C. K. Enders, *Applied Missing Data Analysis*. Guilford Press, 2010.

[28] M. Kuhn and K. Johnson. (2013) Applied predictive modeling. New York, NY. [Online]. Available: http://www.amazon.com/Applied-Predictive-Modeling-Max-Kuhn/dp/1461468485/

[29] B. Butcher and B. J. Smith, "Feature engineering and selection: A practical approach for predictive models," *The American Statistician*, vol. 74, no. 3, pp. 308–309, 2020.

[30] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, 2nd ed. MIT press, 2000.

[31] B. W. Silverman, *Density estimation for statistics and data analysis*, ser. Chapman Hall/CRC monographs on statistics and applied probability. London: Chapman and Hall, 1986. [Online]. Available: https://cds.cern.ch/record/1070306

[32] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 04 1983.

[33] K. A. Bollen, *Structural Equations with Latent Variables*. New York u. a.: Wiley, 1989.

[34] D. BOK, *Higher Education in America*. Princeton University Press, 2013.

[35] R. Arum and J. Roksa, *Academically Adrift: Limited Learning on College Campuses*, 1st ed. University of Chicago Press, 2011.

[36] C. M. Hoxby, "The changing selectivity of american colleges," *Journal of Economic Perspectives*, vol. 23, no. 4, pp. 95–118, December 2009.

[37] V. Tinto, *Leaving College: Rethinking the Causes and Cures of Student Attrition*, 2nd ed. University of Chicago Press, 1993.

[38] J. E. King, “Crucial choices: How students’ financial decisions affect their academic success,” American Council on Education, Center for Policy Analysis, Tech. Rep., 2002.

[39] S. Dynarski and J. E., Scott–Clayton, “The cost of complexity in federal student aid: Lessons from optimal tax theory and behavioral economics,” *National Tax Journal*, vol. 59, no. 2, pp. 319–56, 2006.

[40] V. Tinto, *Completing College: Rethinking Institutional Action*. University of Chicago Press, 2012.

[41] C. Goldin, L. F. Katz, and I. Kuziemko, “The homecoming of american college women: The reversal of the college gender gap,” *Journal of Economic Perspectives*, vol. 20, no. 4, pp. 133–156, December 2006.

[42] D. E. Heller, *The Policy Shift in State Financial Aid Programs*. Dordrecht: Springer Netherlands, 2002, pp. 221–261.

[43] L. W. Perna, *Studying College Access And Choice: A Proposed Conceptual Model*. Dordrecht: Springer Netherlands, 2006, pp. 99–157.

[44] S. L. Thomas and L. W. Perna, *The Opportunity Agenda: A Reexamination of Postsecondary Reward and Opportunity*. Dordrecht: Springer Netherlands, 2004, pp. 43–84.

[45] J. Bound, M. F. Lovenheim, and S. Turner, “Why have college completion rates declined? an analysis of changing student preparation and collegiate resources,” *American Economic Journal: Applied Economics*, vol. 2, no. 3, pp. 129–57, July 2010.

[46] L. DeAngelo, R. Franke, S. Hurtado, J. H. Pryor, and S. Tran, *Completing College: Assessing Graduation Rates at Four-Year Institutions*. Higher Education Research Institute, 2012.

[47] L. Horn and C. D. Carroll, “Nontraditional undergraduates: Trends in enrollment from 1986 to 1992 and persistence and attainment among 1989-90 beginning postsecondary students,” U.S. Department of Education, Office of Educational Research and Improvement, Washington, D.C., Microform, 1996.

[48] J. F. Ryan, “The relationship between institutional expenditures and degree attainment at baccalaureate colleges,” *Research in Higher Education*, vol. 45, no. 2, pp. 97–114, 2004.

[49] G. Kuh, T. Cruce, R. Shoup, J. Kinzie, and R. Gonyea, “Unmasking the effects of student engagement on first-year college grades and persistence,” *Journal of Higher Education*, vol. 79, no. 5, pp. 540–563, Sep. 2008.

[50] C. Adelman, “The toolbox revisited: Paths to degree completion from high school through college,” Office of Vocational and Adult Education, U.S. Department of Education, Washington, D.C., 2006, pdf. [Online]. Available: https://www.loc.gov/item/2006372524/

[51] J. K. Drake, “The role of academic advising in student retention and persistence,” *About Campus*, vol. 16, no. 3, pp. 8–12, 2011.

[52] S. Goldrick-Rab and F. T. Pfeffer, “Beyond access: Explaining socioeconomic differences in college transfer,” *Sociology of Education*, vol. 82, no. 2, pp. 101–125, 2009.

[53] J. L. Stephan, J. E. Rosenbaum, and A. E. Person, “Stratification in college entry and completion,” *Social Science Research*, vol. 38, no. 3, pp. 572–593, 2009.

[54] Complete College America, “Time is the enemy: The surprising truth about why today’s college students aren’t graduating... and what needs to change,” Complete College America, United States, Numerical/Quantitative Data; Reports - Evaluative, 2011, sponsored by Bill and Melinda Gates Foundation; Carnegie Corporation of New York; Ford Foundation; Lumina Foundation for Education; W.K. Kellogg Foundation.

[55] W. G. Tierney and J. R. Sablan, “Examining college readiness,” *American Behavioral Scientist*, vol. 58, no. 8, pp. 943–946, 2014.

[56] D. Hosmer, S. Lemeshow, and R. Sturdivant, *Applied Logistic Regression*, ser. Wiley Series in Probability and Statistics. Wiley, 2013. [Online]. Available: https://books.google.com/books?id=64JYAwAAQBAJ

[57] J. H. Friedman, “Greedy function approximation: A gradient boosting machine.” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189 – 1232, 2001.

[58] J. Pearl, *Causality*, ser. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2009.

[59] G. Ferenstein and B. Hershbein, “How important are high school courses to college performance? less than you might think,” *Brookings*, Jul 2016.

[60] E. Bettinger, A. Boatman, and B. Long, “Student supports: developmental education and other academic programs,” *Future Child*, vol. 23, no. 1, pp. 93–115, 2013.

[61] B. T. Long, “Addressing the academic barriers to higher education,” Brookings, Jun 2014.

[62] D. Xu and S. S. Jaggars, “The impact of online learning on students’ course outcomes: Evidence from a large community and technical college system,” *Economics of Education Review*, vol. 37, no. C, pp. 46–57, 2013.