UTAH’S PATHWAYS TO HIGHER EDUCATION: 
A CRITICAL, QUANTITATIVE ANALYSIS 

by 
Tyler McDaniel 

A Senior Honors Thesis Submitted to the Faculty of 
The University of Utah 
In Partial Fulfillment of the Requirements for the 
Honors Degree in Bachelor of Science 

In 
The Department of Mathematics 

Approved: 

______________________________  ________________________________
Braxton Osting, PhD              Peter Trapa 
Thesis Faculty Supervisor        Chair, Department of Mathematics 

______________________________  ________________________________
Alla Borisyuk, PhD               Sylvia Torti, PhD 
Honors Faculty Advisor           Dean, Honors College
Acknowledgments

Special Thanks to:

Braxton Osting
Erin Castro
Jeremy Franklin
Erica Orians
Thom Gourley

Unless otherwise indicated, data for this research were accessible through Utah’s state longitudinal data system database administered by the Utah Data Alliance (UDA), which includes data supplied by UDA partners and the StudentTracker service from the National Student Clearinghouse. This research including the methods, results, and conclusions neither necessarily reflect the views nor are endorsed by the UDA partners. All errors are the responsibility of the author.
Abstract. This work uses linear and nonlinear models in order to predict student success and pathways in higher education in the state of Utah. Postsecondary Grade Point Average is used as a metric for success in higher education. Pathways are identified using clustering analyses, which group observations according to various distance measures. Eleven institutional pathways were found for the 2008 cohort of Utah high school graduates. Analyzing enrollment data from 2000–2013, four major trends emerged as distinct patterns of post-secondary attendance (or lack thereof) for the cohort: no attendance, early peaking attendance, late peaking attendance, and completion. Underrepresented Racial Minority (URM) students, low income students and geographically mobile students were overrepresented in the group of No-Goers. After identifying pathways, a Random Forest (RF) algorithm was used in order to predict which pathway a student might take based on high school characteristics. Important variables in the random forest prediction include High School GPA, college courses taken in High School, ACT scores, High School proportion of Low Income students, and High School proportion of URM students. Linear models also showed that demographic variables such as race, mobility, and Low Income status, in addition to Pell Grant status carry significant predictive weight for Postsecondary GPA. Attending schools with high proportions of Low Income or URM students adversely affected students’ likelihood of succeeding in college. Pell Grants may have the ability to partially compensate for the disadvantage faced by low income students; Pell Grant eligibility and reception were strongly associated with increased grade point average and retention. The RF algorithm also predicted college attainment correctly at a higher rate than the logistic model. The RF algorithm outperformed Ordinary Least Squares regression models and random coefficients hierarchical linear models in predicting student Grade Point Averages from a test set. The RF algorithm had smaller mean squared error than linear models, as well as smaller absolute medians and quartile values. Due to its accuracy, simplicity and intuitive nature, RF analysis is recommended for future critical quantitative research in higher education. Substantially, the author advocates for an expansion of the Pell Grant program in order to combat the significant postsecondary disadvantages faced by high achieving low income students. Mobility is identified as an underresearched area with high potential impact in postsecondary attainment. Addressing inequities in education pathways, particularly the lasting effects of inferior schools which low income and URM students often attend, will be necessary in the development of the United States as a free, democratic society. Only through critical, equitable, education can members of society liberate themselves from various forms of oppression and reach their full humanity.
## Contents

1. Introduction 1  
   1.1. School and Society 1  
   1.2. Social Climate 3  
   1.3. Critical Theory 4  
   1.4. The Critical Quantitative Possibility 6  
   1.5. Research Contribution 8  
2. Data 9  
   2.1. Source, Structure 9  
   2.2. Cleaning and Organizing 10  
   2.3. Missing Data 11  
3. Dimensionality Reduction and Data Exploration 13  
   3.1. Overview of PCA 13  
   3.2. Exploratory PCA 14  
   3.3. PCA with Restructured Data 16  
4. Clustering Pathways 18  
   4.1. Pathway Context 18  
   4.2. K-Means Clustering 20  
   4.3. K-Medoids Clustering 24  
   4.4. Academic Success Clustering 26  
5. Predicting Pathways and Success 29  
   5.1. Random Forest Background 29  
   5.2. Specifications 31  
   5.3. Variable Importance 34  
   5.4. Partial Plots 34  
   5.5. Pathway Estimation 35  
   5.6. College Attainment Estimation 36  
   5.7. Academic Success Estimation 37  
   5.8. GPA Estimation 38  
6. Model Comparisons for Predicted Student Success 40  
   6.1. Ordinary Least Squares 40  
   6.2. Random Coefficients Model 41  
   6.3. Logistic Model 42  
   6.4. Results 43  
7. Conclusions & Discussion 45  
   7.1. Conclusions 45  
   7.2. Critique 47  
   7.3. High School GPA vs. Standardized Test Scores 48  
   7.4. URM Access 49  
   7.5. Mobility 49  
   7.6. Low Income Access & Pell Grants 50  
   7.7. Critical Contribution 52  
References 54  
Appendix A. R Code 58  
   A.1. Data 58  
   A.2. Dimensionality Reduction and Data Exploration 59
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.3.</td>
<td>Clustering Pathways</td>
<td>60</td>
</tr>
<tr>
<td>A.4.</td>
<td>Predicting Pathways and Success</td>
<td>61</td>
</tr>
<tr>
<td>A.5.</td>
<td>Model Comparisons for Predicted Student Success</td>
<td>61</td>
</tr>
</tbody>
</table>
1. Introduction

Equal access to every level of education is a prerequisite for any society which claims to be equitable. Such access is not always realized, and quality research on educational progression is necessary in order to identify disparities and point to potential solutions. Studied with a commitment to improving parity in education pathways, educational access is measured using a variety of statistical methods, which are useful in studying large and complex social systems. A major goal of this work is to contribute to quantitative critical education scholarship, a research model outlined by Frances Stage (2007). By using new and underutilized algorithms such as RF, quantitative critical scholarship is advanced. Prior to addressing the specifications of the theoretical framework, the researcher explores a broader perspective on the purpose of education and role of critical theory. Beginning with a broad view on the importance of education in a democratic society, the Introduction moves discourse towards the role of education researchers, specifically quantitative critical scholars. Last, the author describes the anticipated impact of the research performed, and makes recommendations for future quantitative critical education research.

Our national discussions of education, our cultural commonplaces about schooling, are pretty much devoid of two themes I think are central to an egalitarian philosophy of education: a robust and nuanced model of intelligence and achievement that affirms the varied richness of human ability, and a foundational commitment to equal opportunity to develop that ability. -Mike Rose, Educator

1.1. School and Society.

The shared responsibility of education is central to any democratic society. The public school, often the manifestation of a child’s first interactions with the state, must be a paragon of society’s values. If the public is committed to a society in which hard work and merit, rather than chance and privilege, determine one’s station in life, then it is essential that their schools be guided by such principles. If the public is committed to scientific investigation, critical research, and the search for truth, then these ideals must be embodied, not only in public schools, but in analyses of these schools.

Social injustices, exploitation, and malfeasance inhibit the creation of an equitable, humanizing society. Using a Freirean (1970) ideology, education is viewed as the only process through which these injustices can be overcome. Despite vast injustices, which threaten the fabric of society, the author is optimistic about the possibilities that emanate from an educated populace. An unjust society is not viewed as an inevitability from which there is no escape, but rather as a limiting situation, which can be transformed through critical and educated resistance (Freire, 1970). As more people gain access to critical education, they will be better equipped to navigate important social issues, and will have the ability to work towards a more complete, humanizing society.
In the 1970’s, Paulo Freire popularized the notion of critical pedagogy. According to Freire, education has the power to liberate students through engendering critical consciousness. This praxis, referred to as conscientization, allows students to gain critical understandings of their social world as it relates to systems of power and oppression. Freire (1970) argued that while schools are often thought to provide students with opportunities, they could in fact do the opposite by teaching students to accept and be complacent with maintaining extant social inequities. One of Freire’s central ideas is that people are not complete creatures, but rather projects, and it is only through conscientization that we may become more fully human. While one’s humanity can be attacked by injustices, exploitation and oppression, it can also be affirmed through conscious resistance learned through education. In the process of oppression, both the oppressor and the oppressed are dehumanized, and therefore it is necessary for every individual to gain conscientization in order to reject dehumanization and recognize their and others’ full humanity.

Freire believes that the oppressed must be responsible for the democratization of society. Inherently, this belief affirms the existence of the oppressed, and acknowledges their inconsequential role in liberation. The oppressor may also play a vital role in transforming society, but only when, in an act of love, he places himself in the situation of the oppressed (Freire, 1970). Through such acts of love, performed by both the oppressed and oppressors, with the motive of freeing themselves from extant social constraints, individuals are humanized. In a research context, the author of this study wishes to present an interpretation which also affirms the humanity of the oppressed. Critical pedagogy asserts that marginalized students, specifically low-income children of Color, are most likely to change the world (Duncan-Andrade, 2007). The critical education of all groups is vital to the improvement of the democratic society; however, this project places additional emphasis on the education of marginalized youth in an effort to spur social change.

Educator Mike Rose 2014 states that public commitment to education is the cornerstone of a free society. Freire (1970) describes freedom as the insatiable quest for human completion. Each believe that freedom is achieved through a liberatory praxis centered around democratic education. Freire likens the process of conscientization to new human being born: not oppressor, nor oppressed, but an individual actively seeking freedom. In order to fully recognize societal freedom, critical education is required. However, freedom is not guaranteed, even when an educated populace exists. Oppression can be self-inflicting: through the inversion of social praxis, the conditions which one believes to be true are absorbed back into social reality. Freire describes this process as “making real oppression more oppressive still by adding to it the realization of oppression.” It is only through deep reflection of society by which one can reveal the true nature of its oppressive conditions. Such thought is not sufficient in procuring freedom, but when critical thought is combined with action, liberation becomes a possibility. Although this type of education is necessary for the progression of democracy, the state may not be keen on fueling such critical thought. James Baldwin 1963 suggests that societies often prefer a docile populace which follows rules. However, Baldwin asserts that such a populace is sure to result in the downfall of that society. The only hope is in responsible, critically thinking individuals, who are willing to do the work necessary to transform society.
Through critical education one can find empowerment, critical consciousness, and liberation from material and psychological oppressions. An educated populace can economically and socially advance their communities. Although the benefits of a sound education extend far beyond the vocational domain, the primary reason for going to school is often to advance one’s station in life. Education, specifically higher education, can provide the tools necessary for class mobility. In 2014, those with Bachelor’s degrees earned 66% more than those with High School diplomas. For each consequent education level attained, median incomes increased significantly (US Department of Education, 2016). Those who study education are becoming increasingly interested in the success of post-secondary institutions. In 2006, the Secretary of Education’s Commission Report on the Future of Education included post-secondary education for the first time, in addition to the normal reports on kindergarten through grade 12. While a college education is increasingly important in the global marketplace, state policies and practices are often inefficient at and in fact discriminatory in – funnelling well-qualified students into higher education (Kirst and Venezia, 2004).

1.2. Social Climate.

Racial gaps in educational opportunity are detrimental to society at large (Chambers, 2009). The inferior pipelines to higher education and STEM fields which underrepresented racial minority (URM) students are dealt inevitably hurt U.S. competitiveness in a global market, a problem exacerbated by the growing populations of Non-White U.S. citizens (Hurtado, 2007). The public education system in the United States is beset with inequities that follow race and class lines. The resegregation of Black and Latino/a public school students, combined with the lack of resources in high-poverty, high-racial minority school districts, have contributed to persistent achievement gaps (Wald and Losen, 2003). Nationally, the Black student resides in a classrooms where 64% of their classmates are low-income (Frankenberg and Orfield, 2012). Race and class inequities such as unequal access, underrepresentation in STEM fields, dissimilar retention efforts or disparate degree attainment, continue to plague the mission of higher education (Bensimon and Bishop, 2012). In 1963, James Baldwin stated that, in reference to the great potential of a marginalized Black child, “if this country does not find a way to use that energy, it will be destroyed by that energy.”

Public schools’ failings often corresponds to criminal activity: Baldwin’s words have been reaffirmed by modern examination of the school-to-prison pipeline, a term used to describe the process by which poor students and students of Color are funneled from classrooms into jail cells. The racial disparities present in disciplinary action rates and the juvenile criminal justice system are so similar that researchers were compelled to connect the dots (Wald and Losen, 2003). The racial climate in the American education system contains inequalities at every level, resulting in systemic failures to educate underrepresented Children of Color.

This project concentrates on Utah’s system of higher education, which replicates national structures of power and access in many regards. In Utah, women attain degrees at higher rates than men. In 2015, 45% of women and 34% of men graduated within 1.5 times the published time for their degree. In the same time period, 54% of women and
45% of men graduated with a Bachelor’s degree in 6 years or less (Utah System of Higher Education, 2015). For the 2008 cohort, Utah institutions graduated 48.7% of women and 41.1% of men. However, some populations remain drastically under-served. Less than 20% of Black and 26.5% of American Indian women graduated, rates that are lower than the 1996 cohort rates for each of these groups (National Center for Education Statistics, 2016). Despite significant increases in the population of underrepresented racial minority citizens, racial diversity at selective, public universities has declined, largely as a result of bans on affirmative action policy (Garces and Cogburn, 2015).

Although higher education in Utah is often similar to higher education nationally, Utah is unique in its large missionary population. Unlike other states, Utah’s citizens are predominately Mormons, or members of the Church of Jesus Christ of Latter Day Saints (LDS) (Huerta and Flemmer, 2005). Men in the LDS Church are strongly encouraged to serve a two-year mission. In 2012, the current President of the LDS Church, Thomas Monson, stated that “we encourage all young men who are worthy and who are physically able and mentally capable to respond to the call to serve. Many young women also serve, but they are not under the same mandate to serve as are the young men.” Previous studies on completion in the state of Utah stress the importance of separating missionaries from dropouts, as many LDS missionaries return to school and complete degrees (Bliss, Webb, and Andre, 2012).

1.3. Critical Theory.

Critical theory is unabashedly political. That is ... critical theory is sustained by an interest in the emancipation of all forms of oppression, as well as by a commitment to freedom, happiness, and a rational ordering of society. Although we may disagree on what these ideas mean, the premise of critical theory- that its purpose is to transform society- is worth heeding when we seek to offer critique via quantitative data. - Benjamin Baez, Educator

Critical scholarship must be placed in a historical and political context. Originating from scholars of the Frankfurt School, critical theory focuses on identifying latent power structures and oppressions, and often manifests in efforts to change existing hierarchies. The work of critical theorists includes changing the methods used to interpret society as well as changing society itself. Critical research is inherently dangerous, as the information which is found and disseminated by critical scholars often upends existing structures of institutional truth (Kincheloe, McLaren, and Steinberg, 2012).

Although quantitative work is prevalent in educational studies, much of the critical research has been qualitative or ethnographic. However, the type of methodology (qualitative, quantitative, etc.) may not be as important as the ways in which researchers propose that individuals reflect, judge, and change society (Baez, 2007). Previous work contends that quantitative and qualitative practices can exist disjunctively, but not conjunctively (Howe, 2009). Qualitative research is associated with interpretivism, which relies on individual experiences and perspectives, rejecting notions of objectivity in the social sciences (Mack, 2010). Quantitative research is associated with positivism, or objective social science research. One might question whether any research can be truly
objective. Freire states that objectivity does not occur by chance, but rather as the product of human action (1970). A growing number of education scholars believe that positivism has no place being aligned with quantitative methods (Howe, 2009). Oppressive empiricism occurs in education when evaluation of education structures is performed without the acknowledgment of group struggles, distribution of resources, and/or current manifestations of oppression.

Education scholars often exclude race altogether, use race as a predictive variable, or compare other races to whites while failing to critically examine racial power structures (Bensimon and Bishop, 2012). For this reason, objectivity is not a goal of this quantitative study. In this work, data are examined using critical interpretivism: methodology and results are discussed with reference to existing social power structures and educational inequities. In this work, for example, the researchers may find that “Underrepresented Racial Minority status is a negative predictor of retention.” In the case of such a finding, the quantitative critical theorist might add “institutions of higher education in Utah are failing to retain Underrepresented Racial Minority students.” In essence, the critical scholar begins with an understanding of racial stratification and the role of that institutions have in creating and maintaining such stratification (Bensimon and Bishop, 2012).

In order to present a critical view of the education system, achievement gaps are alternatively viewed as opportunity gaps, in order to shift the focus from outputs to inputs. In this work, the implication of the achievement gap: that white students achieve more as a result of harder work and greater ability is interrogated (Chambers, 2009). Individual work ethic is important, but a “pull yourselves up by the bootstraps” mentality, without acknowledgement of unequal support structures, is no basis for education policy (Rose, 2014). For these reasons, the term opportunity gap will be used in the majority of this work to discuss disparities in educational success. In doing so, the goal is not to dismiss notions of achievement: as a quantitative researcher, the author believes that accurately measuring variations of human achievement is possible and necessary for advancing curricula. However, national discussions on education have been dominated by the achievement gap model, which places the responsibility of change entirely on the individual student. In this work, it is suggested that the school system also must change in order to adequately instill intelligence and aptitude in its pupils.

One might wonder whether beginning research with assumptions around demographic inequities is inherently biased. This work seeks to shed the guise of objectivity from higher education studies. It is impossible for a researcher to evaluate educational systems from outside the field of education, meaning that true objectivity cannot be obtained. Further, an overwhelming amount of scholars make assumptions about the absence of race, class, or gender disparities prior to engaging with their research, which is perhaps a more severe fallacy than noting the presence of these disparities, due to the immense amount of literature documenting such inequities (Bensimon and Bishop, 2012).

The purpose of this work is to open dialogues around equity and evaluation in higher education. Rather than dictating the way in which one views the world, critical theory lays a map with which one can explore a research question (Kincheloe, McLaren, and
Instead of verifying current models, critical research seeks to challenge and improve on academic understandings of power, privilege and oppression (Stage, 2007). According to Kincheloe and McLaren (2002), critical research is guided by the following principles:

1. Critical research seeks to uncover power relations, which are central to individual and group actions in maintaining or assailing status quos.
2. Critical research examines the social and political forces which prevent individuals and groups from making significant decisions autonomously.
3. The Marxist understanding that economic forces dominate life experiences is rejected, in favor of theory which recognizes racial, gender, and sexual domination, as well as others.
4. Instrumental and technical rationality are rejected; instead of obsessing over procedures, critical scholars focus on the humanistic value of their research.
5. Critical research embraces psychoanalysis as a tool for explaining the complexity of desires which contribute to oppressive structures.
6. Power is understood through Gramsci’s framework of hegemony, which emphasizes the role of psychological domination used by institutions such as media, school, family, and church.
7. Ideology cannot be separated from hegemony, and hegemonic domination is supported by a heterogeneous conglomeration of individuals, each of whom may have nuanced or even countercultural views on various aspects of dominant ideology.
8. Critical research recognizes that language is never neutral; instead discursive practices are mercurial, and can be used to reinforce or challenge power.
9. Culture is salient in power and domination, and therefore critical research must be able to link cultural artifacts to the political economy, and understand the effects of cultural messages on distinct race, class, gender, and sexual groups.
10. Critical research examines the process and impact of cultural pedagogy.

Kincheloe and McLaren’s tenets are often used as a foundation for critical theory (Bensimon and Bishop, 2012; Stage, 2007). One might notice that many of the tenets, particularly (4), seem to be antithetical to quantitative research. The Frankfurt scholars, along with Kincheloe and McLaren and most critical scholars since, envisioned critical theory as exclusively qualitative. However, in the past decade the possibility of a new field of critical thought has emerged.

1.4. The Critical Quantitative Possibility.

Critical education theory has previously been associated with qualitative studies, which typically focus on presenting alternative social narratives in order to represent the experience of marginalized individuals (Stage, 2007). The most important stage of the critical research process is widely considered the interpretation of results; according to the hermeneutical tradition, there is no objectivity, only interpretation (Kincheloe and McLaren, 2002). Many qualitative critical theorists view quantitative work as reductive in nature (Stage, 2007). However, qualitative critical theorists may underestimate the importance of interpretation around statistical results. Statisticians often caution that
their methods are not representations of objective reality, but rather a lens through which one can view data. Leo Brieman and Adele Cutler, the authors of the Random Forest (RF) website, the primary tool of analysis used in this project, offer the following:

RF is an example of a tool that is useful in doing analyses of scientific data. But the cleverest algorithms are no substitute for human intelligence and knowledge of the data in the problem. Take the output of random forests not as absolute truth, but as smart computer generated guesses that may be helpful in leading to a deeper understanding of the problem.

Critical theory, as well as quantitative methods such as random forest, may benefit current understandings of equity and power in education. Both critical approaches and quantitative methods can advance the field of education studies by measuring inequities and challenging oppressive empiricism (2007). Frances Stage defines the role of the quantitative critical theorist as twofold: (1) The quantitative critical theorist will investigate equity of the social world using data. In the modern, data-rich era, large-scale analysis will become increasingly accessible, expanding the possibilities of the quantitative critical field. (2) The quantitative critical theorist will interrogate the usage of current empirical models in educational studies, in an effort to better represent marginalized or oppressed groups (Stage, 2007). Advances in statistical methodology are both frequent and useful for the quantitative critical scholar. The random forest algorithm used in this work was invented in 2001 (Breiman), and continues to be improved upon. The goals of the quantitative critical theorist, namely documenting inequities and challenging methods for representing those inequities, are increasingly viable and important due to surges in information and advances in statistical methodology.

The research question examined in this study comes directly from Stage’s model. First, access to higher education is examined using statewide Utah System of Higher Education data. Of particular interest is the relationship between Pell Grant eligibility and college pathways, the predictive viability and social implications of grade point averages and standardized test scores as measures of achievement, and access to highly esteemed Science, Technology, Engineering and Mathematics fields. Second, the Random Forest model is proposed as an alternative to linear and logistic regression models. As a non-linear, decision tree based, ensemble predictor, the random forest model is structurally dissimilar from typical linear regression models used in education literature.

By assessing the access landscape of higher education in Utah and evaluating the uses of various predictive models, this work answers pertinent questions in critical education scholarship. In 2012, the Association for the Study of Higher Education (ASHE) Institutes on Equity and Critical Policy Analysis identified four pressing needs within higher education. These included (1) promoting the study of racial equity in higher education among young, tenure-track professors, (2) expanding analytic methods used in higher education research in order to include critical scholarship, (3) encouraging policy centers to address racial stratification in higher education, and (4) diversifying pipelines toward non-governmental policy work (Bensimon and Bishop, 2012). This work directly addresses (2) and (3) by proposing new quantitative methods for critically studying higher
education and making policy recommendations for increasing racial diversity in postsecondary access.

1.5. **Research Contribution.**

The goal of this work is to make two contributions, substantial and methodological, to the study of higher education access. Substantially, this work will further current knowledge on access to postsecondary education systems for Utah students of various identities. Methodologically, this work seeks to compare quantitative models for predicting student success in higher education. Both substantial and methodological results are interpreted critically, in an effort to further understandings of access in higher education as well as quantitative critical scholarship.

This work adds to a growing body of quantitative research on higher education access. Previous researchers have used techniques such as Hierarchical Generalized Linear Modeling (Hurtado, 2007; Raudenbush and Bryk, 2002), Multi-Level Modeling (You and Nguyen, 2012), Structural Equation Modeling (Zajacova, Lynch, and Espenshade, 2005), and Random Forest analysis (Hardman, Paucar-Caceres, and Fielding, 2013) in assessing educational progression. Many of these studies have shed light on issues of limited educational access which plague poor communities and communities of color, in addition to impeding the mission of equitable higher education. However, the author’s understanding is that these studies have yet to use Random Forest, a reputedly accurate predictor (Friedman, Hastie, and Tibshirani, 2001) as a tool to predict higher education success and describe equity. This work is driven by the lack of overlap between critical work and Random Forest analyses of higher education.

There is little consistency in the definitions of college pathways. This work seeks to establish data-driven classifiers for higher education pathways. Unlike previous studies, which categorize pathways based on enrollment (Tomkowicz and Bushnik, 2003), enrollment type and selectivity (You and Nguyen, 2012), outcomes of enrollment (Karen, 2002), participation in research (Hurtado, 2007) or graduation (Camara and Echternacht, 2000; Geiser and Santelices, 2007), this work uses a more holistic clustering method, which utilizes time of enrollment, type of enrollment, and institutional characteristics.

In varying academic and public sectors such as finance, biology, and chemistry, Random Forest models are frequently utilized and often regarded as optimal predictors. Despite the RF model’s simplicity and the intuitive nature in which it represents real events, the RF model has not been utilized as much in the field of education. This work seeks to posit the Random Forest algorithm as a viable model for studying equity of access in higher education.

Following Frances Stage’s model of critical quantitative inquiry, the goals of this work are to interrogate both the equity of Utah’s education pathways and the methods which are generally used to study higher education pathways. (1) Focusing on demographic factors such as race, class, gender, and mobility, how equitable are Utah’s pathways to higher education? (2) To what extent can current critical quantitative models of student
success in higher education be improved upon by the use of Random Forest algorithms?

The author finds that identity and geography present barriers to equal education access for many students. While the effects of race and class are well-documented, mobility, school characteristics, and district characteristics also open or restrict pathways for many students. Perhaps the most pressing inequities in higher education access are that of low-income students, and that of low income districts and largely URM schools. Methodologically, RF predictors performed better than logistic and linear models in estimating college attainment and success. The model is intuitive and gives measures of importance for all variables included. Additionally, the RF model allows one to estimate classifications such as the eleven higher education pathways identified by cluster analysis. The use of Random Forest modeling is encouraged in educational assessment and critical research.

Outline. Data organization and imputations are covered in section 2. Sections 3-5 focus on the methodology and results of Principal Components Analysis, Cluster Analysis, and Random Forest Analysis, respectively. Section 6 describes models which were used as comparisons for RF predictions. Section 7 includes results, critique, and conclusions, focusing again on the broader purpose of education in society and the importance of studying education systems empirically.

2. Data

2.1. Source, Structure.

Data were obtained from the Utah System of Higher Education. Data include information on 43,947 students from the 2008 cohort of Utah high school graduates. Not every student necessarily graduated high school, but all were recorded in a Utah high school as part of this cohort. Demographic information includes school district, school, gender, race, low income status, mobility, English Learner status, migrant status, and special education status. Gender was denoted as binary (male/female). Race was divided into ten categories: Caucasian, White not of Hispanic Origin, Black, Asian, Hispanic or Latino, American Indian, Pacific Islander, Multiple Race, and missing. The author acknowledges that the gender and race categories available are not exhaustive, and do not represent every identity of interest. Even so, analysis of the information available was deemed useful. Low Income is indicated for students who qualify for the National School Lunch Program (free or reduced price lunch) or have been identified as economically disadvantaged on another measure during their final year of high school enrollment. Mobile indicates whether a student has attended the same high school for the entirety of that student’s final year of enrollment. Migrant is indicated if the student has been identified as the child of migratory agricultural workers. Special education is indicated if students participated in special education during their final year of high school enrollment. English language learner indicates whether a student participated in a Limited English Program during their final year of high school. Student achievement/recievement information includes Advanced Placement (AP) test scores, ACT test scores, High
Figure 1. A summary of pathways to higher education for students in the state of Utah, the number of students enrolled in various institutions is indicated (students may transfer, and thus be counted in multiple institutions).

School GPA, and cumulative postsecondary GPA. ACT test scores include reading, English, mathematics, science and composite results. Postsecondary GPA is disaggregated by semester. Additional postsecondary information such as Pell Grant eligibility, Pell Grant reception, and course of study (IPEDS definition) were also provided for each semester a student was enrolled in an institution of higher education.

Data were originally structured in four sets, according to student ID, semester, degree, and test. In order to discuss the variety of influences on student success provided, data were combined such that each of the 43,947 students pertained to one row of 186 variables. Variables included all the demographic, testing, and academic summary information previously mentioned. Even so, this grouping of data was not able to capture all the nuances of the semester-level matrix. In order to compensate for this lack of information, several variables were created in order to summarize the students’ higher education experiences. Information such as Cumulative GPA, Earliest Enrollment, and College Semesters in High School were included in order to quantify some of the richness and complexity of postsecondary learning experiences. Many variables were included in this set but ultimately not pertinent to the research question.

In Figure 1, one can see a visual representation of the data used to measure college pathways. The complexity of using pathways to describe college access can be understood, as many students attend college at different times, take breaks, and graduate on different schedules. Even so, there is a basic structure to the institutions and opportunities that are accessible for students.

2.2. Cleaning and Organizing.

STEM students were identified based on the National Center for Education Statistics definition of STEM students, which includes any student who has participated in at least
one semester of a STEM major. Further information on this definition can be found in section 5.4. Thus, a dummy variable was generated indicating whether students had a record of participation in a STEM major or not.

A dummy variable indicating whether students attained college or not was generated based on the 28,878 unique Person ID values from college enrollment data. A second dummy variable was generated to indicate only those who had attained college at some point after 2008. College attendance in high school was then used as a predictor of college attainment post-2008.

Cumulative Undergraduate GPA information was available for every semester that students recorded in institutions of higher education. A time period variable was created, which translated school start dates into numeric values. Then most recent semesters were identified for each student, and Cumulative GPA from these semesters were used. Additionally, a dummy variable was generated which identified semesters recorded prior to Summer 2008 as High School enrollments. These enrollments were counted in order to create a variable displaying the number of college semesters a student recorded in High School.

2.3. Missing Data.

Several items in the dataframe were missing crucial components such as district, school, demographic information and high school GPA. Observations which were missing information from all of the following: district, school, gender, race, migrant status, mobile status, english learner status, and special education status (n = 2644) were removed using listwise deletion. This technique refers to the removal of entire rows of data with missing values. Although listwise deletion is typically not recommended in the case of missing data, there was little advantage in keeping these cases, which had no useful information. Additionally, cases in which the individual had no High School GPA (n = 56) were removed using listwise deletion. Because n is small, in this case 0.15% of the data, such removal is not problematic. However, the High School GPA could also have been estimated, as was done for ACT scores and Pell Grant eligibility.

Missing data is a common problem in education studies. Many students were missing ACT scores, likely because they did not take the test. Fully 44.3% of the 28,878 students who had a record of higher education were missing ACT scores. Test scores can be an important predictor of academic success, and in order to use this predictor it is necessary to estimate the missing values. There is reason to believe that ACT test scores are not missing at random; students who do not have ACT scores may be academically different than their peers. Missing data are classified as nonignorable, as the probability of missing data may depend on the value of the data. In Table 1, one can observe that mean High School GPA for the students who have ACT scores recorded is 3.387, while the mean High School GPA for students who do not have ACT scores is 2.405. Therefore, it is assumed that the differences in academic fortitude that may change the probability of missing ACT scores can be attributed to changes in GPA and other variables. The
<table>
<thead>
<tr>
<th>Six Number Summaries: HS GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Min</td>
</tr>
<tr>
<td>--------------------------------</td>
</tr>
<tr>
<td>Students with ACT Scores</td>
</tr>
<tr>
<td>Students without ACT Scores</td>
</tr>
</tbody>
</table>

**Table 1.** Students without ACT scores are academically different than those with ACT scores.

Missing values are assumed to be *missing at random* (MAR).

Multiple Imputation (MI) was performed using the Amelia package in R to estimate ACT scores and Pell Grant status. One of the assumptions of MI is that the data follow a multivariate normal distribution. If \( X \) is an \( n \times q \) matrix with missing and observed portions, and \( \theta = (\mu, \Sigma) \) are mean and covariance parameters, then:

\[
X \sim N_q(\mu, \Sigma)
\]

Although this method works best with multivariate data, MI can also work well with non–normal data (Allison, 2001). The multivariate assumption tends to perform fairly well compared to more complex models, even when data follow other distributions. Another assumption for MI is that the data are MAR. Notice that this is different from *missing completely at random* (MCAR), which assumes that missing data have no dependence on any values, missing or random. Instead, the missingness of data is dependent on the observed variables present. In this case, it is assumed that the missing status of ACT and Pell Grant Status Variables are dependent on other observed data, such as High School GPA and demographic indicators. Given that \( M \) indicates the missing data and \( X^{obs} \) are observed data:

\[
p(M|X) = p(M|X^{obs})
\]

This equation is expanded:

\[
p(M, X^{obs}|\theta) = p(M|\theta)p(X^{obs}|\theta)
\]

The MI algorithm is only able to compute observed parameters, so the following Likelihood is used:

\[
L(\theta|X^{obs}) \propto p(X^{obs}|\theta)
\]

Using the Law of Iterated Expectations, this can be rewritten:

\[
p(X^{obs}|\theta) = \int p(X|\theta)dX^{mis}
\]

Where \( X^{mis} \) are the actual missing data values. At this point, assuming a flat prior on \( \theta \), one can obtain:
The Expectation–Maximization (EM) algorithm has been used in previous education studies to estimate missing values (Hurtado et al., 2008). The Amelia package uses EM in combination with bootstrapping in order to find the mode of (2.6) and estimate $\theta$. Then $X^{mis}$ is predicted based on $X^{obs}$ and $\theta$ using a linear regression. For more on Amelia MI specifications, see (Honaker, King, Blackwell, et al., 2011). The range of ACT scores is restricted to 0–36, and the range of Pell Grant statuses is restricted to 0–1. This is done by discarding any estimated values that appear outside of this range. Because the estimated parameters are discrete, the values were rounded to the nearest integer. The imputations were performed five times, and estimates generally become more accurate every time the imputations are performed. Therefore, the fifth imputation was used in order to replace missing data.

3. Dimensionality Reduction and Data Exploration

3.1. Overview of PCA.

Principal Components Analysis (PCA) is used in preliminary analysis in order to better understand the shape and components of the data. PCA was also helpful in choosing variables of importance and opening discussion around underlying factors. PCA is an appropriate decision tool for reducing the number variables needed from a dataset in order to analyze its most important components. For large databases, PCA analysis can reduce the dimensionality while maintaining the maximum possible variation within the data (Everitt and Hothorn, 2011). PCA components are often used as axes for representing data trends in two dimensions.

The goal of PCA, given an $n \times q$ matrix $X$, is to describe a set of correlated variables $X^T = (x_1, \ldots, x_q)$ by a set of uncorrelated components $Y^T = (y_1, \ldots, y_q)$, where each $y_i$ value is a linear combination of $x$ values $y_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{iq}x_q$. The first component, $y_1$, accounts for the most variation possible among the linear combinations $Y$. The next component, $y_2$, is chosen such that it represents the most remaining variation in the linear combinations, given it is orthogonal to the first component. One might note that by increasing the coefficients $a_{i1}, a_{i2}, \ldots, a_{iq}$, a researcher could increase the variance (or variation accounted for) of $y_i$. This problem is solved by adding the constraint that $a_i^T a_i = 1$. The variance of component $y_i$ is defined as $a_i^T S a_i$, where $S$ is the sample variance matrix of $X$. It turns out that $a_i$ is the eigenvector of $S$, such that $a_1$ corresponds to the largest eigenvalue, $a_2$ to the second largest eigenvalue, and so on. By constraining $a_i^T a_j = 0$, where $i \neq j$, one is able to to ensure that $y_i$ and $y_j$ are uncorrelated, which is a necessary condition for the components to be orthogonal.

Notice that the total variance of $q$ eigenvalues is the total variance of the original variables, $\sum_{i=1}^{q} \lambda_i^2 = s_1^2 + s_2^2 + \cdots + s_q^2 = \sum_{i=1}^{q} s_i^2 = \text{trace}(S)$, where $S$ is the sample
covariance matrix of $X$. Therefore, we can represent the proportion of total variance that each component accounts for by

\[ P_i = \frac{\lambda_i}{\text{trace}(S)} \]

The first $m$ components account for $P(m)$, a proportion, of the variation of $X$. This is defined:

\[ P(m) = \frac{\sum_{i=1}^{m} \lambda_i}{\text{trace}(S)} \]

Because PCA depends on correlated variables and the covariance matrix $S$, it is generally recommended to use continuous or ordinal values. However, there was significant motivation to include factors such as race, gender and low-income status in exploratory analysis, as these variables pertained to the research question. Therefore, factor variables were coded as binary variables, which may be used in PCA when there is no alternative (Vyas and Kumaranayake, 2006).

Because there were large differences in the ranges of variables (ACT scores range from 0 to 36, binary demographic information range from 0 to 1), the correlation matrix, $R$ was used in place of the covariance matrix $S$. The correlation matrix simply represents a standardized version of $S$. This substitution is recommended in the case of highly disparate covariances (Everitt and Hothorn, 2011).

### 3.2. Exploratory PCA.

Initially, principal components were constructed using the 42 numeric variables in the dataset, including demographics, test scores, semesters spent in different institutions, and graduation information. PCA performed poorly in terms of describing the data; at most $\sim 35\%$ of the variance could be explained by the first five components. In Figure 2, one can observe the percentage of variance explained by the first $i$ components of three groups of data.

One might expect that the cumulative proportion of variance explained ($P_i^{(m)} = \frac{\sum_{i=1}^{m} \lambda_i}{\text{trace}(S)}$) would be generally smaller for binary variables. For this reason, PCA was attempted on subsets of data which included no binary variables.

Because the amount of information available for each student was contingent on their college access and graduation data, PCA is performed on three distinct sets: All Students, College-Going Students, and College Graduates. In Figure 3, one may observe the edges in the PCA plot for all students, as well as college-going students. This feature is likely a result of the large proportion of students who did not go to college, and thus had 0s for many variables (semesters, AP credits, etc.). For college graduates, two clusters are visible within the PCA plot, likely corresponding to 4-year graduates and 2-year
Figure 2. Scree Diagrams of PCA applied to three groups of observations. The component variance remains fairly high in all models, indicating that the PCA leaves much of the variance in the data unexplained. This is largely attributed to binary variables.

Figure 3. Plots of the first two components of three different PCA results.

One may wonder how the presence of demographic variables would affect PCA results. The addition of binary variables (gender, underrepresented racial minority status, low income status, pell-grant status, mobility status, special education status) did not significantly effect the first two principal components. However, more components were necessary in order to explain a reasonable proportion of the variance. Typically, PCA analyses without binary variables explained ≈ 70% of the variance in five components, while PCA analyses with binary variables explained only ≈ 50% of the variance in five components.

For College-Going students (Table 3.2), the first principal component can be interpreted as ACT score. The second component is interpreted as a students' retention (semesters and graduation). The third component is interpreted as motivation (degree
intent and four year colleges), the fourth component is largely a student’s participation in 2-year college, and the fifth component is a student’s enrollment in a 2-year private institution.

For College Graduates (3, the first principal component is once again interpreted as ACT score. The second component can be interpreted as traditional graduation path (4-year public university, degree intent, graduation). The third component is interpreted as a student’s liberal arts talent (humanities ACT scores and GPA). The fourth component can be viewed as college enrollment (semesters, enrolled in HS), the fifth as college readiness (AP scores), and the sixth and seventh components correspond to the type of post-secondary institution attended.

In every PCA model, the ACT scores had the largest absolute loadings in the first component. Math and science ACT scores were consistently higher than reading and English ACT scores. This result affirms the education literature that finds composite and math scores to be the largest predictors of college success (Noble and Sawyer, 2002; Camara and Echternacht, 2000).

3.3. PCA with Restructured Data.

Because the goal of this work is to document college pathways, data were restructured in order to capture student enrollment timelines. Every observation corresponds with 164 variables, indicating institution type (2 year public, 2 year private, 4 year public, 4 year private) for each semester in which enrollment was possible. Enrollment type was

<table>
<thead>
<tr>
<th>Variable</th>
<th>Comp. 1</th>
<th>Comp. 2</th>
<th>Comp. 3</th>
<th>Comp. 4</th>
<th>Comp. 5</th>
<th>Comp. 6</th>
<th>Comp. 7</th>
<th>Comp. 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Year Private</td>
<td></td>
<td>-0.261</td>
<td>0.215</td>
<td></td>
<td>-0.602</td>
<td>0.157</td>
<td>-0.345</td>
<td>-0.299</td>
</tr>
<tr>
<td>2 Year Public</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.259</td>
<td></td>
</tr>
<tr>
<td>4 Year Private</td>
<td>0.230</td>
<td>-0.280</td>
<td>-0.451</td>
<td>0.182</td>
<td></td>
<td></td>
<td>0.197</td>
<td></td>
</tr>
<tr>
<td>4 Year Public</td>
<td>0.230</td>
<td>-0.280</td>
<td>-0.451</td>
<td>0.182</td>
<td></td>
<td></td>
<td>0.197</td>
<td></td>
</tr>
<tr>
<td>Semesters</td>
<td>0.247</td>
<td>-0.488</td>
<td>-0.173</td>
<td></td>
<td>0.133</td>
<td></td>
<td>0.177</td>
<td>-0.179</td>
</tr>
<tr>
<td>College HS</td>
<td>0.162</td>
<td>-0.304</td>
<td>0.204</td>
<td>-0.274</td>
<td></td>
<td></td>
<td>0.571</td>
<td></td>
</tr>
<tr>
<td>IPEDS Level</td>
<td>0.217</td>
<td>-0.369</td>
<td>0.101</td>
<td></td>
<td>0.160</td>
<td>-0.291</td>
<td>-0.243</td>
<td></td>
</tr>
<tr>
<td>ACT Read</td>
<td>0.331</td>
<td>0.244</td>
<td></td>
<td>-0.262</td>
<td></td>
<td>0.303</td>
<td>-0.298</td>
<td></td>
</tr>
<tr>
<td>ACT Math</td>
<td>0.370</td>
<td>0.227</td>
<td></td>
<td>-0.235</td>
<td></td>
<td>0.313</td>
<td>-0.306</td>
<td></td>
</tr>
<tr>
<td>ACT English</td>
<td>0.340</td>
<td>0.215</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACT Composite</td>
<td>0.379</td>
<td>0.230</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACT Science</td>
<td>0.304</td>
<td>0.250</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degree Intent</td>
<td>0.141</td>
<td>-0.201</td>
<td>-0.540</td>
<td></td>
<td></td>
<td></td>
<td>0.100</td>
<td>0.128</td>
</tr>
<tr>
<td>UGrad GPA</td>
<td>0.213</td>
<td>-0.188</td>
<td>0.279</td>
<td>0.299</td>
<td>-0.146</td>
<td></td>
<td>-0.359</td>
<td>0.399</td>
</tr>
<tr>
<td>HS GPA</td>
<td>0.255</td>
<td>-0.170</td>
<td>0.208</td>
<td>0.303</td>
<td>-0.141</td>
<td></td>
<td></td>
<td>0.404</td>
</tr>
<tr>
<td>AP Credits</td>
<td>0.200</td>
<td>-0.144</td>
<td>0.284</td>
<td>0.112</td>
<td>-0.508</td>
<td></td>
<td>-0.114</td>
<td>-0.637</td>
</tr>
</tbody>
</table>

Table 2. Component Loadings for PCA performed on college going student data
Table 3. Component Loadings for PCA performed on college graduate data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Comp. 1</th>
<th>Comp. 2</th>
<th>Comp. 3</th>
<th>Comp. 4</th>
<th>Comp. 5</th>
<th>Comp. 6</th>
<th>Comp. 7</th>
<th>Comp. 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Year Private</td>
<td>0.180</td>
<td>-0.288</td>
<td>-0.120</td>
<td>0.300</td>
<td>0.518</td>
<td>0.996</td>
<td>0.116</td>
<td>0.278</td>
</tr>
<tr>
<td>4 Year Private</td>
<td>-0.231</td>
<td>0.524</td>
<td>-0.229</td>
<td>0.139</td>
<td>-0.190</td>
<td>0.792</td>
<td>0.278</td>
<td></td>
</tr>
<tr>
<td>Semesters</td>
<td>-0.105</td>
<td>0.267</td>
<td>-0.168</td>
<td>0.535</td>
<td>-0.287</td>
<td>0.149</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPEDS Level</td>
<td>-0.222</td>
<td>0.403</td>
<td>0.159</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACT Read</td>
<td>-0.279</td>
<td>-0.417</td>
<td>0.274</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACT Math</td>
<td>-0.350</td>
<td>-0.183</td>
<td>-0.157</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACT English</td>
<td>-0.283</td>
<td>-0.403</td>
<td>0.282</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACT Composite</td>
<td>-0.368</td>
<td>-0.174</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACT Science</td>
<td>-0.341</td>
<td>-0.175</td>
<td>-0.216</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degree Intent</td>
<td>-0.167</td>
<td>0.425</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UGrad GPA</td>
<td>-0.230</td>
<td>-0.137</td>
<td>0.395</td>
<td>0.239</td>
<td>0.312</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS GPA</td>
<td>-0.223</td>
<td>0.273</td>
<td>0.183</td>
<td>0.158</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP Credits</td>
<td>-0.263</td>
<td>0.216</td>
<td>-0.244</td>
<td>-0.468</td>
<td>-0.169</td>
<td>0.218</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

denoted on a 0-7 point ordinal scale (0=Not Enrolled, 1=Withdrawn, 2=Less than Part Time, 3=Part Time, 4=Half Time, 5=Unknown, 6=Leave of Absence, 7=Full Time). Students listed as Unknown were often attending small schools outside the state. Students listed as taking a Leave of Absence often are completing a mission before enrolling in school Full Time. While the ordering of enrollment type could affect components analysis results, the Not Enrolled, Part Time and Full Time groups were much larger than the others, and there is ample separation between these three groups.

For the restructured data, the first two components are interpreted as Private Attendance without Graduation, and 2 Year Institution Attendance, respectively. As seen in Figure 4 first two components only account for $\approx 34\%$ of the total variance, which is not surprising as the data are ordinal and contain many zeros. The PCA plot shows many striations in the data, although it is unclear what these represent. There are no groupings of observations.

PCA is useful in identifying major components and reducing the amount of information necessary in describing a matrix. The first components provide effective axes for plotting data and displaying variation. Additionally, some conclusions about variable importance can be made based on loadings, and important variables can be used in further models. While PCA is useful in reducing dimensionality and describing data, PCA does not allow for clear interpretations of relationships among data. The following sections delve into the interactions between variables.
4. Clustering Pathways

4.1. Pathway Context.

The classification of de facto education pathways is vital to the study of higher education and a major tenet of this work. Clustering analysis is a useful tool for classifying data in which “true” or “correct” groups are unknown. Using statistical methods in order to classify data was desirable due to the immense information available and the variety of ways in which scholars have defined college pathways. K-means clustering allowed the researcher to group data according to several variables. Previous researchers often categorize college pathways by enrollment time, enrollment type, institution type, and academic success; cluster analysis allows one to group observations according to any combination of these variables. By clustering the data according to distance between observations, the researcher was able to analyze all information available with empirical methods.

The number of students pursuing some type of postsecondary education has risen dramatically over the past two decades. Total undergraduate enrollment in postsecondary institutions increased 31% from 2000-2014, from 13.2 million students to 17.3 million students. The US Department of Education projects that total undergraduate enrollment will be 19.8 million in 2025. Much of the increases in postsecondary enrollment can be attributed to Underrepresented Racial Minority students: Hispanic undergraduate enrollment doubled and Black enrollment increased 57% between 2000 and 2014, while White enrollment only increased 7% during this time period (US Department of Education, 2016). As colleges continue to diversify, it will be important to account for

---

**Figure 4.** Scree Diagram and plot of the first two Principal Components of the restructured data
the differences in opportunity andreceipt that exist at each level of higher education. Community colleges, which serve a diverse population of students, often provide transfer possibilities to high SES or White students that are not available for students of color or poor students (DOWD, 2007). Alicia Dowd argues that education scholars must transcend the access “saga”, which relies on preconceived notions of education pipelines such as the 2 Year Public to 4 Year Institution transition, and instead focus on student outcomes (2007).

Student outcomes can be difficult to measure. The college access data were limited to a five-year time-frame, which is not enough time for many students to attain a degree, especially those who enroll in 4-year institutions and serve two-year Latter Day Saint missions, which are popular in Utah. However, many missionary students will return to school and complete degrees. Many Utah students took college courses in high school, which is fundamentally different than attending a post-secondary institution, a difference that should be accounted for. Thus, there is significant motivation to separate this group from those who dropped out or never enrolled, even though the “outcome” in each case may be no degree. While much of the previous research around college access has been focused on attainment, it is increasingly important to measure success in higher education in terms of type of access, retention, and degree attainment. In order to capture a full understanding of the higher education landscape, it is necessary to identify the nuanced post-secondary educational pathways that students take (You and Nguyen, 2012).

While many students enroll in higher education in Utah, few graduate within 5 years. Less than 70% of Utah’s 2008 high school graduating cohort enrolled in postsecondary education. By 2013, only 18% of the original cohort had attained a postsecondary degree of any sort. Even so, the problems that this work addresses transcend any particular stage of the education process. Due to the diversity in college enrollment and graduation patterns, a greater understanding of educational pathways is sought. Clustering analysis is used in order to group data according to patterns in college attendance. Unlike previous research, which uses binary outcomes of enrollment (Karen, 2002), participation in research (Hurtado, 2007) or graduation (Camara and Echternacht, 2000; Geiser and Santelices, 2007), this work identifies eleven distinct college pathways, based on time of enrollment, type of enrollment, and institutional characteristics.

A number of clustering techniques can be used in order to group data points into various categories. Clustering is desirable for grouping college pathways due to the wide variety of methods that previous researchers have used to describe pathways. Canadian researchers Tomkowic and Bushnik (2003) define three distinct routes to college: the Right-Awayer (enrolls in post-secondary institution within 12 months of high school graduation), the Delayer (enrolls in post-secondary institution more than 12 months after high school graduation), and the No-Goer (has no record of post-secondary enrollment). In a recent publication, The Condition of Education 2016, the U.S. Department of Education offers various ways to describe of higher education pathways, based on attainment, institution type, enrollment type, and institution selectivity. Attainment outcomes include a high school diploma or equivalency certificate, an associate’s degree, a bachelor’s degree, or a master’s degree. Institution types include 4 year Public, 4 year
Private nonprofit, 4 year Private for profit, 2 year Public, 2 year Private nonprofit, and 2 year Private for profit. Enrollment type refers to full-time or part-time. Institutional Selectivity is categorized according to the percentage of students admitted. Researchers You and Nguyen (2012) categorize higher education pathways according to institutional type and selectivity, as well as students’ course-taking patterns, in a multi-level analysis. While all the aforementioned methods are valid in grouping student outcomes, a pathway classification which accounted for all or many variables is desirable. By using clustering analysis, the researchers allow distance between data points to determine the best classifications of educational pathways.

4.2. K-Means Clustering.

K-means clustering is one method used in order to separate \( n \) individuals in a data set into \( k \) groups (\( G_1, G_2, G_3, \ldots G_k \)). The choice of \( k \) is bestowed upon the researcher, although methods exist for selecting \( k \). Generally, the researcher is aware of a range of partitions that would be viable given the research context. In terms of college pathways, based on previous work, anywhere from 2 to 24 post-secondary pathways would be feasible (one might combine several combinations of variables, e.g. institution type, enrollment type and time-frame, to create 12-24 unique pathways). One can measure the effectiveness of a K-means grouping using the Within-Group Sum of Squares, defined as:

\[
WGSS = \sum_{j=1}^{q} \sum_{l=1}^{k} \sum_{i \in G_l} (x_{ij} - \bar{x}_j^{(l)})^2
\]

where \( X \) is an \( nxq \) matrix and \( \bar{x}_j^{(l)} = \frac{1}{n_l} \sum_{i \in G_l} x_{ij} \) is the mean of values in group \( i \) on variable \( j \) (Everitt & Hothorn, 2011). Given a desired number of clusters \( k \), a K-means algorithm will seek to partition the data in a manner such that \( WGSS \) is minimized. If it were possible to inspect every possible partition, this would be simple. However, even the fastest computers are not equipped to compute \( WGSS \) for every possible partition. Instead, computers seek to find local minima using the following method:

1. Assign points to \( k \) clusters. This can be done using an initializing method.
2. Calculate the change in clustering criterion that occurs when moving each point to a different cluster.
3. Make change that results in greatest improvement of clustering criterion. \( WGSS \) is minimized by \( G(i) = \text{argmin}_{1 \leq l \leq k} ||x_i - \bar{x}_j^{(l)}||^2 \)
4. Repeat steps (2) and (3) until assignments do not change.

Because K-Means clustering uses Euclidean distance as a dissimilarity measure, it has the tendency to produce spherical clusters, even when the data are truly clustered according to some other geometric pattern. Additionally, K-Means clustering is not scale-invariant (Everitt and Hothorn, 2011). However, due to the binary nature of the institution and degree type data available, scaling problems are not an issue. The spherical nature of K-Means clustering is addressed in further detail in Section 4.3.

Using the kmeans command in R, K-Means clustering classifies students according to institution type, enrollment type, time period, and graduation, the correct number of
clusters was unclear. Because the K-Means algorithm finds local minima in the sum of distances from cluster centers, the clustering assignments can vary each time it is performed. In order to obtain a “good” clustering, WGSS is minimized over 20 trials and plotted. One may note that in Figure 5, while a clear elbow is indistinguishable, eleven appears to be a reasonable number of clusters.

Clustering appears to be coherent: in Figure 6, groups of certain colors are easily identifiable. However, the effectiveness of the clustering is unclear from this perspective. Because of the high dimensionality of the data, it is difficult to know whether this is a “good” clustering or not. Also, it is impossible to interpret clusters simply by inspecting the PCA plot.

Table 4 provides a tabular representation of the clustering of student pathways in higher education in Utah. One may notice the disproportionality of Cluster 10, the No-Goers, which represents nearly 57% of all 43,947 student observations. While this is desirable in the sense that all of these students have little or no college attendance records, such uneven grouping can cause problems for machine learning estimation techniques (Liaw & Wiener, 2002), which will be addressed in the Random Forest section. Otherwise, pathways are relatively uniform, with 2 Year Public and 4 Year Public clusters being larger than 4 Year Private and Transfer clusters. Although 4 Year Private schools were included in the data as a viable institution type, not enough students ($n < 200$) chose this path for the K-means algorithm to identify it as a cluster.
Figure 6. K-means clusters plotted along the first two principal components, colored according to its cluster label.

Figure 7 provides additional insights on the K-means pathway clustering, displaying each cluster according to its institution type and proportion of students enrolled. The shape of the points also shows the type of enrollment (Full Time, Part Time, etc.). Further, degree attainment rates are shown, with circles conveying more than half of the cohort received a particular degree. Institution type is a key distinguisher in the eleven pathways. Students are largely grouped according to their institution type. Additionally, four distinct patterns of enrollment are salient: No-Goers, Peak Early, Peak Late, and Completers. This pattern appears in nearly every major institution type. There is not as much distinction around enrollment type; most students are either Full Time or Not Enrolled.

It is important to differentiate between the students that attended school early and those who attended school in later years. Students who attended early, but are no longer enrolled may be more less likely to complete degrees than students who attend school later. Clusters 9, 3 and 8, the Peak Late clusters, had the three lowest proportions of women. Because the students who peaked late in college enrollments were overwhelmingly male, the author interprets this finding as an identification of missionary pathways. In Figure 8, one can observe that Low Income students, URM students, mobile students, special education students and English language learners are all over-represented in the No-Goers pathway, indicating broken pathways for these populations. This majority of this pathway was comprised of mobile students.
Figure 7. K-Means clusters, mapped according to institution type, enrollment type, and degrees attained. Distinct institutional patterns and groups of No-Goers, Early Peakers, and Late Peakers are evident. Titles are colored according to the cluster colors in Figure 6.
Table 4. K-means clustering groups displayed according to the type of institution and attendance patterns displayed. Four main pathways were discovered in terms of attendance patterns: students either stopped out after high school (never went to college), peaked early (enrolled, then perhaps stopped out), peaked late (perhaps took two year leave of absence, then attended), or (nearly) completed their degrees.


Like K-means clustering, K-medoids clustering is a technique used for classifying data into \( k \) distinct groups \((G_1, G_2, G_3, \ldots G_k)\). Rather than using Euclidean distance between points, the K-medoids clustering method uses Manhattan distance. This method may make more sense when using ordinal data.

K-means clustering relies on Euclidean distance to minimize the \( WGSS \), a process which is performed iteratively until the cluster assignments remain stagnant. Although computationally intensive, K-Medoids clustering offers an alternative which does not rely on purely quantitative data. Because K-Medoids uses Manhattan distance rather than Euclidean, the separation between two classes is measured in integer amounts. This method makes more sense in measuring patterns of enrollment; the distance between two students is essentially the number of dissimilar semesters. An arbitrary method for calculating K-medoids clustering, based on distance, is outlined as follows:

1. Given a cluster \( G \), and an \( nxq \) matrix \( X \), find the observation which minimizes the distance to all other points in the cluster: 

\[
i_l^* = \arg\min_{i \in G_l} \sum_{i' \in G_l} D(x_i, x_{i'}).\]

Then \( x_{i_l^*}, l = 1, 2, 3, \ldots k \) are the current estimates of cluster centers. Using Manhattan Distance, 

\[
D(x_i, x_{i'}) = \sum_j |x_{ij} - x_{i'j}|\]

2. Given the set of cluster centers, minimize the total error by assigning each observation to the closest cluster center: 

\[
G(i) = \arg\min_{1 < l < k} D(x_i, x_{i_l^*})
\]
Figure 8. Clusters are represented according to their demographic characteristics. The *Peak Late* clusters are all predominantly male, while groups such as Low-Income, URM, and Mobile students are overrepresented in the *No-Goers* cluster. Point size is proportional to cluster size, indicating the disproportionate population of the *No-Goers* cluster.

(3) Repeat steps (1) and (2) until the assignments do not change.

The *WGSS* plot for K-Medoids clustering is more difficult to interpret than that of K-Means clustering. The kcca command in the Flexclust package was used in order to separate groups according to “kmedians”. However, R lacked the computational power to find the *WGSS* for just one cluster (the entire data set). It is assumed that two or more clusters are appropriate. In Figure 9, after taking the minimum *WGSS* of 20 trials, increasing the number of clusters sometimes leads to greater *WGSS*. This should not be the case. With more trials, one would be able to plot a non-increasing *WGSS* function. Eleven clusters were chosen in order to directly compare the K-Medoids and K-means clustering methods. K-Medoids clustering also distinguishes the *Early Peakers*, *Late Peakers*, *No-Goers*, and *Completers*. The visual representation of K-Medoids clusters in Figure 14 is similar to that of K-Means clusters in 6, although there may be slightly more overlap between groups. In Figure 11, one notes that K-Medoids clusters are similar to those of K-Means clustering, although there are no Private school clusters. Due to the difficulty in plotting the *MWGSS* for K-Medoids, the uncertainty that the correct number of clusters were chosen, and the similarity with K-Means clustering, cluster demographics are not displayed. K-Means clustering is preferred, although the author believes that with some modification K-Medoids clustering could provide better results.
Figure 9. \textit{MWGSS} (Minimum Within Group Sum of Squares) for 20 trials. Even after 20 trials, the function is not monotone. For this reason, the “correct” number of clusters is difficult to ascertain.

4.4. Academic Success Clustering.

In addition to identifying the institutional and temporal pathways that students commonly take, clustering students based on the level of academic success attained in college was also desirable. Students in Science, Technology, Engineering and Mathematics (STEM) fields generally attain more degrees and persist longer than those in other disciplines (Chen, 2009). Due to the research emphasis of Female (Clark Blickenstaff*, 2005), Low Income (Chen, 2009) and URM (Hurtado et al., 2010; Museus et al., 2011) access to STEM fields, students’ intent and participation in STEM field majors was structured by semester. Cumulative GPA, in addition to STEM intent and participation, were clustered in order to identify student success in higher education. The clustering largely relied on GPA, as the researchers recognize that not all students aspire to study STEM fields. However, because the underrepresentation of women (Beede et al., 2011) and URM students (Hurtado et al., 2010) is widely documented in STEM fields, these variables were added in order to broaden the discussion around equity and access.

Whereas the National Science Foundation (NSF) includes social/behavioral sciences in their definition of STEM fields, most contemporary scholars limit STEM classification to mathematics, natural sciences, engineering, computer sciences, and related fields (Chen and Weko, 2009). The STEM classification used in this work mirrors Chen and Weko’s. Classification of Instructional Program (CIP) code categories of (27) Mathematics, (3) Natural Sciences, (40,41) Physical Sciences, (26) Biological/agricultural sciences,
(14,15) Engineering/engineering technologies, and (11) Computer/information sciences were selected using the National Center for Education Statistics (NCES) Integrated Post-secondary Education Data System (IPEDS). Further information can be found in the footnotes section.

Using the breadth of information available, K-means clustering was applied to post-secondary GPA and STEM data in order to separate successful and unsuccessful students. Unlike the pathway clustering, the WGSS plot in Figure 12 clearly shows that two clusters are the appropriate choice. This was the case every time the model was performed, and therefore the WGSS results were not minimized over 20 trials. Figure 13 displays the data plotted along its first two principal components, colored according to their assigned cluster.

Figure 13 displays the mean GPA and STEM involvement for the two groups identified by K-means clustering. The second cluster clearly has much higher GPA and even STEM participation. Therefore, this group is labeled the academically successful cluster. The first cluster, with lower GPA and STEM participation, is labeled as the academically unsuccessful cluster. The author recognizes that all academic success cannot be measure with grade point averages or STEM participation. Even so, these variables are the best proxies available for measuring student excellence in higher education.
Figure 11. K-Medoids clusters mapped according to enrollment type, institution type, and degrees attained.
5. Predicting Pathways and Success

5.1. Random Forest Background.
Figure 14. The two clusters identified by the K-means algorithm differ drastically in average GPA. Although STEM variation is difficult to notice, the second cluster also averages much higher STEM participation.

College pathways are predicted using an RF decision tree algorithm. Classification and Regression Trees (CART) such as RF are desired due to their intuitiveness, flexibility and ease of use (Golino, Gomes, et al., 2014). The RF model uses random samples of predictor variables to generate decision trees that form forests. Each tree’s prediction is then aggregated in order for the forest to make an estimation. The forest can make estimations for nominal or continuous variables, both of which are utilized in this work. The RF model has been used widely across academic fields in order to assess risk and make predictions. In addition to returning decision trees with split points, the RF model can be used to represent the predictive power of variables included. The RF model is ideal for college access data because it performs well with large datasets and has robust algorithms for dealing with missing data (Hardman, Paucar-Caceres, and Fielding, 2013).

Although many measures of student progression and success are used in education literature, many institutions are unaware of the variables which best predict student progress. In addition to performing well with large datasets, RF models are able to identify variables that are important in predicting outcomes. A growing body of research seeks to understand post-secondary educational pathways through quantititative methods. RF algorithms are increasingly used in order to predict educational attainment or progression (Blanch and Aluja, 2013; Cortez and Silva, 2008; Golino, Gomes, et al., 2014; Hardman, Paucar-Caceres, and Fielding, 2013). However, all of the aforementioned works evaluate educational systems outside the United States, and none investigate post-secondary attainment. The Random Forest model remains relatively unexplored as a tool for describing equity in higher education in the United States.

Although Random Forest Algorithms were developed in statistics (Breiman et al., 1984) and machine learning (Quinlan, 1993), RF methods are being used effectively
to make predictions in fields such as systems biology (Geurts, Irrthum, and Wehenkel, 2009), ADHD diagnosis (Skogli et al., 2013), medicinal chemistry (Naeem, Hylands, and Barlow, 2012), and finance (Khaidem, Saha, and Dey, 2016). In fact, RF models are becoming increasingly popular in consumer finance, where misclassification can be costly. Estimations of consumer finance risk are important both due to the large number of consumers taking loans and the implications of poor modeling, as exemplified by the credit crunch of 2008 (Thomas, 2010). Since the 1950s, banks have been using classification methods in order to produce credit scores, which indicate the level of risk that a bank must assume in order to provide a loan. Banks generally wish to identify two classes of risk: good loans (non-defaulting) and bad loans (defaulting). Due to changes in banking regulations enacted at the Basel III Accord of 2009, which advise banking regulators to provide protections against future economic downturns by changing capital and liquidity rules. In short, the Basel Accords seek to minimize credit risk. Many researchers have used RF algorithms in order to predict credit risk (Brown and Mues, 2012; Iturriaga and Sanz, 2015; Wang et al., 2011). The use of Random Forest algorithms in diverse and complex fields, especially consumer finance where errors are highly costly, illustrates the predictive power of the algorithm.

A growing body of research seeks to understand post-secondary educational pathways through quantitative methods. While progression through higher education has been measured using RF algorithms, the Random Forest model remains relatively unexplored as a tool for describing equity.

Random forests are a popular method for estimating data, due to their simplicity and accuracy. The random forest regression is an ensemble learning algorithm, similar to methods such as bagging and boosting. Each type of ensemble learning can be used in prediction, involving the aggregation of individual learners (such as trees). While boosting typically dominates bagging, random forests perform similarly to boosting and are easier to adjust (Friedman, Hastie, and Tibshirani, 2001). The random forest model can be used in regression or classification problems, using either the average of individual decision tree results for regression, or a weighted vote of the decision tree results for classification (Breiman, 2001).

Random Forest trees are known for being noisy, but have relatively small bias if they are grown adequately deep. While boosting trees adjust in order to correct for bias, RF trees are identically distributed. The expectation of any RF tree is the same as the expectation for a committee of trees, but variance can be reduced by using more trees. Aggregating over many trees has been shown to produce accurate predictions for large and complex data (Friedman, Hastie, and Tibshirani, 2001). An example decision tree can be seen in Figure 15.

5.2. Specifications.

The method for obtaining a random forest predictor is as follows:

(1) Take $K$ bootstrap samples from the training data.
(2) For each bootstrap sample, grow a random forest tree:
- Select \( m_{\text{try}} \) variables at random\(^1\)
- Pick the best variable/split point\(^2\) among the \( m_{\text{try}} \) variables
- Split the node into two daughter nodes.

(3) Repeat step 2 until minimum node size \( n_{\text{min}} \) is reached\(^3\)

(4) Predict new data by aggregating predictions from sample trees

Let an ensemble of \( K \) trees be represented as \( \{T_k\}_{1}^{K} \). For a point \( x \), a new prediction is generated as follows:

\[
\hat{f}_{rf}^{K}(x) = \frac{1}{K} \sum_{k=1}^{K} T_k(x)
\]

**Regression**: Given that \( \hat{C}_k(x) \) is the class prediction of the \( k^{\text{th}} \) tree, \( \hat{C}_{rf}^{K}(x) \) is the majority vote of the random forest trees(\( \{\hat{C}_k(x)\}_{1}^{K} \))

**Random Forest Notes:**
1. The results turn out to be insensitive to the number of features selected to split each node. Usually, selecting one or two features gives near optimum results. Selecting a small \( m_{\text{try}} \) at random allows each variable a chance to appear in any given decision tree. This may be useful in measuring the effects of variables which would otherwise go unused. Brieman recommends using the default \( m_{\text{try}} \) values of \( \sqrt{p} \) for classification and \( \frac{p}{3} \) for regression, as well as one half the default and twice the default (Liaw and Wiener, 2002). Reducing \( m_{\text{try}} \) will reduce the correlation between any two trees, henceforth reducing the average variance of the ensemble. The correlation between two bootstrapped trees is typically small (less than 0.10) even for large matrices (\( m_{\text{try}} \) up to 100), and variances of bootstrapped trees are not much larger than variances of original variances, meaning the trees are effective in reducing average variance, \( \rho \sigma^2 + \frac{1-\rho}{T} \sigma^2 \) (Friedman, Hastie, and Tibshirani, 2001).

2. The best variable/split point can be found by minimizing the sum of the following error terms with respect to the cut \( S \):

\[
\min \sum_{i:x_{ij} \geq S} (y_i - \hat{y}_i)^2 + \min \sum_{i:x_{ij} \leq S} (y_i - \hat{y}_i)^2
\]

3. The default node size is 1 for classification and 5 for regression (Liaw and Wiener, 2002). By increasing the node size, one can simplify the random forest model.

Given an \( nxq \) matrix \( X \), a bootstrap sample of size \( n \) is simply a random sample \((x_1, x_2, \ldots x_n)\), taken with replacement. The bootstrap sample is typically used in order to generalize a sample based finding to a larger population. Bootstrapping is often used in assessing margin of errors. Rather than sampling the population numerous times, which can be costly and inefficient, one may use a bootstrap to gather information on the population distribution (Singh & Xie, 2008).
Figure 15. Example decision tree for pedagogical purposes. Notice the split points, chosen by minimizing the error on each side (5.2)
Out of Bag (OOB) samples allow one to make conclusions about the effects of individual variables. For each observation $x_j$, the OOB estimate constructs its random forest predictor by averaging only those trees which did not use $x_j$. OOB samples are used to calculate error in classification.

5.3. Variable Importance.

Variable importance is a measure of the increase in OOB with a permutation of a particular variable (Liaw and Wiener, 2002). Brieman and Cutler’s website describe the permutation technique for computing Variable Importance as follows:

1. For each tree in the forest, take the OOB estimates and sum the correct votes cast $c$.

2. Randomly permute the value of variable $m$ in the OOB cases, and recast votes according to these values. Count the correct votes cast, $c_{\text{permute}}$.

3. The average over all trees, $\frac{1}{K} \sum_{k=1}^{K} (c - c_{\text{permute}})$

One commonly used proxy for Variable Importance is Mean Decrease Gini (MDG), which is the average decrease in Gini Impurity for a predictor, across a forest (Nicodemus, 2011). The Gini Impurity is measured using the following method: When a node is split on variable $m$, the Gini Impurity criterion for the descendant nodes must be less than that of the parent node. The sum of decreases in Gini Impurity gives a measure of importance for variable $m$ that is often consistent with the permutation method (Breiman and Cutler, 2014).

Thus, information is gathered on the influence of the variable of interest. In regards to educational equity, there are three types of variables to consider: achievement-based measures (test scores, GPA), institutions (2-year/4-year, public/private) and identity factors (gender, race, class, mobility). If college pathways were equitable, one would expect achievement measures to be the most important predictors for post-secondary success (retention, GPA, graduation).

According to three random forest models, GPA was consistently the most important factor in prediction of college GPA and graduation. Unsurprisingly, institutions were important in terms of the type of degree a student might receive (although not as important as GPA). Gender was consistently more important than all other demographic factors. Linear models support the notion that female gender may be a significant positive predictor of academic success. However, models thus far do not account for persistent gender discrepancies in Science, Technology, Engineering, and Mathematics (STEM) pathways (Beede et al., 2011). Pell grants were fairly important relative to other demographic traits, and linear models indicate that pell grants may have a compensatory effect for the consistent disadvantage that Low Income students face.

5.4. Partial Plots.

Partial plots are obtained using the average values for all other variables, $x_{i\neq C}$ in order to measure the change in prediction that occurs when the variables of interest, $x$, is
manipulated (Liaw and Wiener, 2012). For regression, the change in prediction is simple to measure. For classification, the following is used to measure change in votes:

\[
(5.3) \quad f(x) = \log p_k(x) - \frac{1}{K} \sum_{j=1}^{K} \log p_j(x)
\]

Partial plots provide evidence to the direction of a predictor. Figure 21 displays the partial plots for demographic (gender, race, class, mobility) and recievement (ACT scores, HS GPA) variables in their relation to undergraduate GPA. Partial plot results generally showed the same directional relationship as linear model coefficients.

5.5. Pathway Estimation.

A central focus of this work is the prediction of college pathways based on high school recievement and demographic characteristics. Therefore, using pathways identified by the K-means clustering algorithm, an RF model is produced in order to predict student pathways.

A Random Forest model was used in order to classify the data according to the eleven pathways identified by the K-means Clustering. Data was divided into training and test sets; in this case only 500 observations were used in the test set. The RF model used had specifications of \(m_{try}=4\), and \(K=500\) trees.

The test set confusion matrix in Table 5.5 gives the researcher another view into the accuracy of the RF model. In this case, the model is extremely accurate when predicting cluster 10, the No-Goers, and not very accurate for any other groups. This is fairly common when one group is much larger than the others; an extremely simple algorithm would simply estimate that every student belonged to cluster 10. This algorithm would misclassify students about 43% of the time- so the RF model is much better. One method for combating this bias is to assign a mandatory share of the predictions to a small group (Liaw and Wiener, 2002).

Error rates for Random Forest models typically stabilize as \(K\), the number of trees, gets large (Friedman, Hastie, and Tibshirani, 2001). In Figure 16, one can see that the error rates for the different classes remain nearly constant beyond 200 trees. Even though large forests are not necessary for RF estimation to be accurate, they do provide more stable Variable Importance Measures (VIM). Brieman often uses 1,000 or even 50,000 trees in order to assess variable importance (Breiman and Cutler, 2014). In this model, variable importance is measured at \(K=500\). High School GPA appears to be extremely important in pathway classification. While individual demographic characteristics such as URM and Low Income were not very important, school factors such as URM proportion and Low Income proportion were highly important in the forest. District characteristics were not as important as school characteristics, but still more important than individual level demographics. Pell grant eligibility and reception, as well as the number of college courses taken in High School, carried predictive power. PCA results suggested that Math
36

Confusion Matrix: Test Data
Prediction of Cluster Membership

<table>
<thead>
<tr>
<th>Actual</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>15</td>
<td>2</td>
<td>1</td>
<td>40</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>23</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>11</td>
<td>15</td>
<td>4</td>
<td>79</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>8</td>
<td>4</td>
<td>8</td>
<td>14</td>
<td>12</td>
<td>114</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>14</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>11</td>
<td>115</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>12</td>
<td>10</td>
<td>4</td>
<td>170</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>10</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>28</td>
<td>22</td>
<td>3</td>
<td>70</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>68</td>
<td>8</td>
<td>46</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>64</td>
<td>69</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>15</td>
<td>17</td>
<td>1989</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>9</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>191</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>134</td>
<td>3</td>
<td>33</td>
<td>0.8325</td>
</tr>
</tbody>
</table>

Table 5. The test set total error rate is 40.08%

Figure 16. Error Rates and Variable Importance for Pathway Estimation. The error rate for Cluster 10, the No-Goers, is much lower than the rest. Clusters 4 and 8, both of which represent 4-year Private school pathways, also have relatively low error rates.

and Science ACT scores contained the most variation of the standardized test scores, however the RF model points to English and Reading being the most important subjects.

5.6. College Attainment Estimation.

After predicting specific college pathways, the author turned to the broader question of college attainment. In addition to being a simpler prediction, this type of prediction
can be compared to other models in order to evaluate the effectiveness of RF. This estimation also serves as the first step in predicting student GPA, which can only be assessed for college-going students.

In Figure 17, one can see that the total error rate for college attainment estimation is less than a quarter. Also interesting is the fact that the Variable Importance Plot shows nearly the same results as Figure 16. One exception is that AP scores above 3 are less predictive. This finding was not expected, as those who take AP courses would generally be expected to have college aspirations and plans. The binary variable indicating whether students enrolled in college courses while in high school was much more important in this estimation than in pathway estimation, suggesting that while AP scores may not influence a student’s postsecondary plans dramatically, enrollment in college level courses may.

5.7. Academic Success Estimation.

Academic success was predicted using the Cluster Analysis results from GPA and STEM data. K-means Clustering indicated the clear separation of two groups, interpreted as successful and unsuccessful students; these groups were predicted as an alternative to using an RF regression to predict GPA. Similar to financial RF estimations, college admissions personnel may be interested in predicting Completers and Dropouts, as completion rates are often used for a benchmark of success among universities (DeAngelico et al., 2011). Because few students had vast STEM records, the academic success clustering relied heavily on GPA. The researcher is essentially estimating GPA using a classification method. The comparison of classification and regression methods for predicting student success may be useful for college admissions researchers and policymakers.

In Figure 18, one can observe that the variable importance measures in academic success estimation are similar to those in the Postsecondary GPA regression prediction.
Figure 18. Error rates and Variable Importance for predicting academic success. The total error rate is 26.4% model. It appears the Successful Cluster are much easier to predict than the Unsuccessful Cluster. While the demands of the researcher ultimately determine which prediction method is best-suited, the classification model appears to perform favorably.

5.8. GPA Estimation.

In addition to estimating the educational pathways of Utah students, success in the form of Grade Point Average is predicted. GPA is commonly used as a metric for student success (Geiser and Santelices, 2007; Noble and Sawyer, 2004), and is not affected by leaves of absence. Predicting GPA also allows the researcher to compare RF estimation directly with other common forms of student success estimation.

In order to assess the accuracy of random forest predictions, it is appropriate to use a sample of training data in order to predict a test set (Breiman, 2001). A random sample of 3,740 data points (approximately 10% of the observations) was selected as test data, and the remainder of the data were used as the training set.

By comparing Figure 19 to Figures 16, 17, and 18, it is clear that importance estimators were similar for all three models. Student success/recievement characteristics such as High School GPA, ACT scores, and college course taking are top predictors in each case. In the RF estimation of GPA, Pell Eligibility was much more important than Pell Grants themselves. Reasons for this finding are unclear, although this suggests that it may be useful to further study students who are eligible for Pell Grants but do not receive them.

Note that while the shape of the partial plots and relative scale of the y-axis are important, the y-axis values are not meaningful, as results represent change after integration over all other variables (Liaw, 2009). Interestingly, in Figure 20, GPA predictions declined as many of the ACT scores approached their upper limits. Reasons for this are unclear, although this finding calls into question college admissions offices’ reliance on
**Figure 19.** Error and Variable Importance for Postsecondary GPA Estimation.

**Figure 20.** Partial plots of RF Postsecondary GPA Prediction, in order of variable importance (left to right, top to bottom). Of interest are the bell shapes of many of the ACT curves, as well as the directionality of Pell Grants and school characteristics.
ACT scores as measures of potential success for high achieving students. Pell Grant eligibility and reception are clear positive indicators for Postsecondary GPA. The proportion of URM students in a school district was positively associated with Postsecondary GPA. However, the proportions of URM students and Low Income students in a High School were important and each showed negative trends in relation to Postsecondary GPA. This implies that students from high URM or Low Income schools are less prepared to succeed in higher education than their peers. This finding presents an affront to ideals of equal opportunity and meritocracy in education.

6. Model Comparisons for Predicted Student Success


Two linear models were also used to estimate student success in higher education. The Ordinary Least Squares (OLS) regression method is common tool in educational and sociological data analysis. The Hierarchical Linear Model (HLM) adds new levels of detail to the linear regression. For the HLM model, coefficients estimated at one level become outcomes at the next level (Bryk & Raudenbush, 1986). The multi-level nature of this model allows one to nest students within schools, or schools within districts. This is helpful in segregating the effects of individual characteristics and the effects of group-level characteristics. The Random Effects (RE) model was chosen due to its popularity in the social sciences (Bell & Jones, 2014) and its allowance for estimating group level characteristics, such as school racial composition or district economic composition.

The RF model is compared to both linear models using cross validation. Observations are divided into training and test data, and after using the training data to specify the model, test set values are predicted. Cross validation allows the researcher to compare the accuracy of RF and linear models, in order to make conclusions about the utility of each in predicting higher education success.

Ordinary Least Squares (OLS) is the most common methods used for multivariate analysis (Dismuke & Lindrooth, 2006). OLS is a linear estimator which minimizes the squared residuals in order to estimate new data. The OLS regression id denoted:

\[ Y_{ij} = \beta_0 + \beta_1 X_j + \epsilon_{ij} \]

such that \( \beta \) are coefficients, \( X_j \) represent individual student variables, and \( \epsilon_{ij} \) are error terms.

In Table 6, one can observe from the \( t \) values that High School GPA is the most significant predictor of postsecondary GPA. Every unit increase in High School GPA is associated with nearly a half unit increase in postsecondary GPA. AP scores are also highly significant, although the GPA increase associated with passing AP scores is much less than that of High School GPA. While URM status and Low Income status both have significant negative effects on individual success, Low Income District Mean and URM District Mean appear to have opposite effects on student success in higher education.
| Variable                                | Coef. | Std. Err | t    | p>|t| |
|-----------------------------------------|-------|----------|------|-----|
| (Intercept)***                         | 0.997 | 0.039    | 25.863 | 0.000 |
| Female**                                | 0.029 | 0.012    | 2.474  | 0.013 |
| High School GPA***                     | 0.489 | 0.011    | 45.594 | 0.000 |
| Low Income***                          | -0.074| 0.016    | -4.511 | 0.000 |
| Migrant*                               | 0.513 | 0.294    | 1.742  | 0.081 |
| Special Education                      | -0.035| 0.031    | -1.115 | 0.265 |
| Mobile***                               | 0.039 | 0.014    | 2.797  | 0.005 |
| Limited English                        | -0.011| 0.057    | -0.194 | 0.846 |
| Attend College HS***                   | 0.036 | 0.003    | 11.228 | 0.000 |
| Pell Eligible                          | 0.048 | 0.040    | 1.191  | 0.234 |
| Pell Grant***                          | 0.113 | 0.041    | 2.775  | 0.006 |
| AP Scores > 3***                       | 0.057 | 0.005    | 10.761 | 0.000 |
| URM**                                  | -0.051| 0.020    | -2.588 | 0.010 |
| ACT Read                               | -0.004| 0.004    | -0.971 | 0.332 |
| ACT Math**                             | 0.010 | 0.004    | 2.373  | 0.018 |
| ACT English                            | 0.006 | 0.004    | 1.389  | 0.165 |
| ACT Science                            | -0.003| 0.004    | -0.653 | 0.514 |
| ACT Composite                          | 0.004 | 0.016    | 0.250  | 0.803 |
| STEM***                                | -0.078| 0.015    | -5.180 | 0.000 |
| Low Income District Mean***            | -0.916| 0.127    | -7.211 | 0.000 |
| Low Income School Mean                 | -0.003| 0.117    | -0.022 | 0.983 |
| URM District Mean***                   | 1.038 | 0.138    | 7.506  | 0.000 |
| URM School Mean                        | -0.012| 0.121    | -0.101 | 0.919 |

Table 6. Coefficients for OLS estimation of Postsecondary GPA *==significant at α = 0.1, **==significant at α = 0.05, ***==significant at α = 0.01

This may be the result of differences in rural and urban districts; further investigation in district characteristics could shed light on such an oddity. School effects were not significant. Mobile is a significant positive indicator in the OLS model, although the RF partial plot indicated that Mobility is a negative indicator of postsecondary success. Pell Grant reception is a significant positive indicator of postsecondary success. Interestingly, STEM classification is negatively associated with postsecondary GPA. Researchers from UCLA found that students with aspirations of medical school are nearly 12% less likely to persist in STEM fields (Chang et al., 2014). This is attributed to the highly competitive nature of medical school. Similarly, STEM fields could be more competitive, and therefore more difficult to obtain high levels of success (GPA) than other fields.

6.2. Random Coefficients Model.

There is often motivation to control for groupings in data, such as high schools, which have been shown to significantly influence a child’s educational trajectory (Burtless, 1996). In order to examine potential differences between high schools, one can observe that retention (semesters) and success (GPA) vary by school or district.
| Variable                     | Coef.  | Std. Err | DF    | t     | p>|t| |
|------------------------------|--------|----------|-------|-------|------|
| (Intercept)***               | 0.963  | 0.096    | 18106 | 10.055| 0.000|
| Female                       | 0.013  | 0.012    | 18106 | 1.146 | 0.251|
| High School GPA***           | 0.556  | 0.011    | 18106 | 48.717| 0.000|
| Low Income***                | -0.069 | 0.016    | 18106 | -4.274| 0.000|
| Migrant*                     | 0.560  | 0.290    | 18106 | 1.930 | 0.054|
| Special Education            | -0.025 | 0.031    | 18106 | -0.806| 0.420|
| Mobile                       | 0.018  | 0.016    | 18106 | 1.160 | 0.246|
| Limited English              | -0.044 | 0.056    | 18106 | -0.791| 0.429|
| Attend College HS***         | 0.030  | 0.004    | 18106 | 8.669 | 0.000|
| Pell Eligible                | 0.018  | 0.040    | 18106 | 0.473 | 0.636|
| Pell Grant***                | 0.149  | 0.041    | 18106 | 3.638 | 0.000|
| AP Scores > 3***             | 0.054  | 0.005    | 18106 | 10.109| 0.000|
| URM                          | -0.031 | 0.019    | 18106 | -1.591| 0.112|
| ACT Read                     | -0.003 | 0.004    | 18106 | -0.758| 0.448|
| ACT Math**                   | 0.008  | 0.004    | 18106 | 1.971 | 0.049|
| ACT English                  | 0.007  | 0.004    | 18106 | 1.578 | 0.115|
| ACT Science                  | -0.002 | 0.004    | 18106 | -0.473| 0.637|
| ACT Composite                | 0.001  | 0.015    | 18106 | 0.038 | 0.970|
| STEM***                      | -0.073 | 0.015    | 18106 | -4.894| 0.000|
| Low Income District Mean *    | -0.697 | 0.398    | 52    | -1.752| 0.086|
| Low Income School Mean *     | -0.339 | 0.180    | 103   | -1.877| 0.063|
| URM District Mean *          | 0.334  | 0.445    | 52    | 0.750 | 0.457|
| URM School Mean*             | 0.331  | 0.199    | 103   | 1.665 | 0.099|

Table 7. Coefficients for OLS estimation of Postsecondary GPA *=significant at α = 0.1, **=significant at α = 0.05, ***=significant at α = 0.01

To account for high school-level differences in post-secondary retention and success, one may use a random intercept model, which gives each grouping a separate intercept:

\[ Y_{ij} = (\beta_0 + u_{ij}) + \beta_1 X_j + \epsilon_{ij} \]

where \( i \) are schools, \( j \) are individual students, \( u_{ij} \) are the between group error terms and \( \epsilon_{ij} \) are the within group error terms, such that \( u_i \sim N(0, \sigma_u^2) \); \( \epsilon_{ij} \sim N(0, \sigma^2) \).

In Table 7, the coefficients in the Random Effects Model are predominantly similar to those in the OLS model. However, school and district effects are much less significant. Additionally, Mobility is no longer significant.

6.3. Logistic Model.

The Logistic model is often used in order to estimate binary outcomes. The Logistic Regression assumes that binary outcome variables follow a binomial distribution, and can
be modeled as a function of the independent variables. The variance of this distribution changes as a result of the observation (Cabrera, 1994).

\begin{equation}
E[Y_i = 1|X = x] = P(y_i = 1)
\end{equation}

\begin{equation}
\sigma_i^2 = P(y_i = 1)(1 - P(y_i = 1))
\end{equation}

The Logistic model is typically presented in probabilities, which are useful for interpreting the effect of independent predictor variables on a binary outcome. The intercepts and coefficients are estimated by Maximum Likelihood (ML) estimation, which minimizes the error using prior distributions of these variables (Cabrera, 1994). In the case of modeling college attainment, $Y_i$ represents a student's success (1) or failure (0) to enroll in postsecondary education after Spring 2008. In order to create the confusion matrices in Tables 6.3 and 9, probabilities were rounded to the nearest integer, which were always 0 or 1 due to the constraints of the Logistic model. The RF model performs slightly better than the Logistic model in estimating college attainment.

\begin{equation}
P(Y) = \frac{e^{\beta_0 + \beta_1X_1}}{1 + e^{\beta_0 + \beta_1X_1}}
\end{equation}

6.4. Results.

One of the central goals of this work is to compare Random Forest algorithms to linear estimators. Because linear models are generally used for regression and not classification, the Random Forest GPA estimation is compared to linear GPA estimation. The RF model compares well to both linear and hierarchical linear models. One favorable quality of the RF model is its range of estimation. Because the model is based on aggregated
## Six Number Summaries: GPA Estimation

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1st Quartile</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Quartile</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Data</td>
<td>0.029</td>
<td>2.550</td>
<td>3.133</td>
<td>2.952</td>
<td>3.570</td>
<td>4.000</td>
</tr>
<tr>
<td>Random Forest</td>
<td>1.115</td>
<td>2.573</td>
<td>2.884</td>
<td>2.921</td>
<td>3.280</td>
<td>3.961</td>
</tr>
<tr>
<td>Linear Model</td>
<td>1.346</td>
<td>2.672</td>
<td>2.974</td>
<td>2.933</td>
<td>3.233</td>
<td>4.181</td>
</tr>
<tr>
<td>Hierarchical Linear Model</td>
<td>1.228</td>
<td>2.656</td>
<td>2.979</td>
<td>2.930</td>
<td>3.251</td>
<td>4.274</td>
</tr>
</tbody>
</table>

Table 10. The RF estimator outperforms OLS and HLM models in all prediction metrics except mean and median.

## Six Number Summaries: Error Terms (Predicted GPA–Actual GPA)

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1st Quartile</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Quartile</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>-2.029</td>
<td>-0.440</td>
<td>-0.107</td>
<td>-0.031</td>
<td>0.274</td>
<td>2.749</td>
</tr>
<tr>
<td>Linear Model</td>
<td>-2.316</td>
<td>-0.445</td>
<td>-0.125</td>
<td>-0.019</td>
<td>0.308</td>
<td>2.833</td>
</tr>
<tr>
<td>Hierarchical Linear Model</td>
<td>-2.511</td>
<td>-0.451</td>
<td>-0.122</td>
<td>-0.022</td>
<td>0.312</td>
<td>2.912</td>
</tr>
</tbody>
</table>

Table 11. The RF estimator outperforms OLS and HLM models in all error metrics except mean.

**Figure 21.** Histogram and density plot of test set errors for linear models as well as RF model. Although the mean error for the RF model is slightly larger than the others, clearly the RF model is the more accurate estimator.

decision trees which are calculated from randomly sampled data, the RF model will not estimate GPAs under 0 or above 4.

Table 6.4 provides insight in the type of predictions made by the various models tested. None were good at estimating the minimum value. However, the means and medians
Figure 22. Visualization of Actual and Predicted Values, for Random Forest (left), Ordinary Least Squares (middle), and Hierarchical Linear Modeling (right). One may note the similarities between the estimations for both linear models. The RF model appears to more accurately estimate low performing college students, although it tends to underestimate high performers.

for all model estimations were quite close to the actual values. While RF was slightly outperformed on these measures, it did a better job approximating the 1st and 3rd quartiles, as well as the minimum and maximum. The linear models both predict GPA values above 4, an undesirable quality in this type of prediction model. The RF model fits the shape of the data better, even though measures of centrality are slightly less precise. In Table 6.4, one can observe that the RF model had a larger mean absolute error term in predicting the test data (although the linear models had higher error mean error squared terms), and the quantile values for the RF model error terms were much closer to 0. The median error value in the RF model was -0.107, compared to -0.125 for the linear model and -0.122 for the hierarchical linear model. Visually, one can tell by Figure 21 that the RF model consistently produces estimates that are closer to the actual test set values.

In Figure 9, one can see that the linear models essentially produce two classes of estimates, one for those with a Postsecondary GPA and one for those without. This is not necessarily undesirable, although the RF model does appear to better fit the data. While the linear models are similar predictors, one can observe that the RF model uses a unique prediction algorithm. Both predictors slightly underestimate the GPA for many high achieving students.

7. Conclusions & Discussion

7.1. Conclusions.

This work makes two contributions, methodological and substantial, to critical studies of higher education. Methodologically, statistical tools such as cluster analysis and
Random Forest were used in order to identify and predict student pathways in higher education. Clustering was used to classify students based on patterns of postsecondary enrollment. The RF algorithm is used as a nonlinear predictor of higher education pathways and higher education success. Compared to more commonly used models such as OLS, HLM, and Logistic estimation, the RF algorithm predicted student enrollment and success more accurately. Substantially, demographics, academic measures and policies were considered in order to accurately describe Utah’s pathways to higher education. Factors such as race, gender, low income status, mobile status, school, and district were investigated for equity of access in higher education. Pell Grants were considered as a potential tool for the dismantling of income related barriers to higher education. The utility of academic measures such as GPA, ACT scores and AP scores in the admissions process is interrogated. By providing a summary of methodological and substantial findings, as well as a discussion on implications, the author wishes to further the possibility of an equitable education system.

Principal Components Analysis is used in Section 3 to explore data. Primarily, this study used PCA to reduce the dimensionality of data and observe the variables containing the most variation. In this sense, ACT Scores were the most important variable in all PCA models. Further, ACT Math, Science and Composite scores were consistently the variables with the highest first component loadings, corroborating education literature, which states that math and science scores are particularly important in the college-going process. Alternatively, ACT scores were not nearly so important in the subsequent models predicting college pathways and success. While math scores remained important in RF and linear models, other ACT subjects were not significant. ACT scores were important in RF models, but Reading and English scores were often more important than Math scores.

Random Forest algorithms are recommended for future work in predicting student success and pathways in higher education. The RF method is simple, flexible, logical, and accurate. Critical quantitative studies of higher education could be improved by the accuracy and applicability of RF algorithms. Fundamentally different from linear predictors, the RF model predicts both classification and regression problems with accuracy, outperforming HLM, OLS and Logistic models in student success estimation. Additionally, the RF model is interpretable, with variable importance readily ascertained from its decision trees. Partial plots are useful in deducing the directionality of predictor variables. Although partial plots may be used exclusively, the researcher finds it helpful to validate interpretations of partial plots with linear model coefficients. Using both linear and nonlinear estimators, a greater understanding of demographic effects is achieved.

Certain trends are clear in all models the researcher examines. High School GPA is by far the most important predictor of college attainment, education pathway and student success. ACT test scores are important in RF models and fairly insignificant in linear models, although Math scores are important in both methods. Taking college courses in high school is also a fairly important predictor of postsecondary success. AP test scores are less important. High School GPA, ACT and AP test scores, and high school college enrollment in High School are all positive indicators of postsecondary GPA. Pell grant
reception is a significant positive indicator of postsecondary success in every model. Although this work primarily focuses on the demographic inequities in education pathways, it is necessary to state that demographic factors pertain to a lesser degree of statistical importance than academic factors. Even so, identities clearly matter in this process. Low Income status is consistently the most significant demographic predictor of success in higher education. Low Income status is a negative predictor of success in higher education, as are the proportion of Low Income students in an individual’s school and district.

There are areas in which the various methods show seemingly contradictory results. These areas are discussed as potential topics for future studies. While linear models show Mobility as a positive indicator of postsecondary GPA, RF models illustrate Mobility as a barrier to postsecondary attainment and achievement. Similarly, RF models show that the proportion of URM students in an individual’s school is a negative predictor of postsecondary success, while linear models suggest this is insignificant or a positive indicator. Special Education status and Migrant status, though clearly important identities to study, are generally found to be insignificant in the models examined. The author believes that there may not have been enough data on these groups to draw meaningful conclusions. As with other results, stronger assertions could be made using national data.

7.2. Critique.

The scope of this work largely covered classification and prediction methodology; the discussion around missing data imputation was limited. However, the method by which data are imputed could have potentially drastic effects on the predictive analyses. While Multiple Imputation is a valid method for substituting missing data, other methods such as Maximum Likelihood are commonly used in education (Hurtado et al., 2008). Similar results, using a variety of imputation methods, could bolster the conclusions of this study.

For PCA analysis, the correlation matrix was used in place of the covariance matrix, \( S \). However, mean-centering binary variables may not make much sense. For this reason, it may be worth mean-centering all variables except those that are binary. Because PCA depends on highly correlated variables, one could try using a heterogenous correlation in order to correlate binary variables. In terms of random forests, one could change the classification prediction rules in order to adjust uneven group bias (Liaw and Wiener, 2002). Factor analysis could provide meaningful interpretations of potential underlying factors, such as intelligence or social disadvantage.

This study did not attempt to measure school selectivity within institutional categories. That is, while it is assumed that 4 year private institutions are more selective than 2 year public institutions, measurements of selectivity (acceptance rate, average test scores) were not included in the data. Education literature suggests that more selective colleges may be more effective in retaining and graduating students, although this could be the result of college preparedness (Alon and Tienda, 2005; Light and Strayer, 2000). Future work could further delineate postsecondary pathways according to selectivity of
In future cluster analysis of education pathway data, the use of K-Medoids clustering is recommended. While this study primarily relies on K-Means clustering, the use of Manhattan distance in K-Medoids clustering is a more logical classification approach for this type of data. The K-Medoids clustering algorithm is computationally extensive, and it is possible that the number of trials performed in this study was not adequate for choosing an appropriate number of clusters. This may be true of K-Means clustering as well. Future work could predict clustering of fewer or more pathways.

In this work, all models were run without substantial modification. One asset of the RF model is its ability to perform well even with little adjustment. Even so, Brieman (2002) recommends running the model with $m_{try}$ as half the default value and twice the default value. Additionally, Brieman recommends using more than the default 500 trees when creating variable importance plots. Different values for $m_{try}$ and $K$ (number of trees) were attempted for early RF models, resulting in little differences in accuracy or importance. However, with some persistence the models could likely be improved upon. Similarly, the weighting and variables included in linear and Logistic models could likely be modified in order to create slightly more accurate predictions. Additionally, further techniques, such as Structural Equation Modeling, could be compared to RF. While this could be interesting, the author believes that comparing all models in simple forms contributes to the parity of comparison. The author believes it is unlikely that, even with substantial modification, other models would have outperformed RF.

7.3. **High School GPA vs. Standardized Test Scores.**

Freshmen grades in college are the most common outcome measure for predictive studies on student success in higher education. However, many predictors, such as High School GPA, may be more important in predicting the later years in a student’s college career (Geiser and Santelices, 2007).

High school GPA typically predicts first year college GPA more accurately than standardized test scores, although there is some debate about the utility of each measure in predicting subsequent achievement (Sawyer, 2013). Admissions test scores may be more useful in predicting higher levels of achievement at more selective institutions (Sawyer, 2013; Noble and Sawyer, 2004). In addition to being a better predictor of initial student success, High School GPA has been shown to contain less bias against URM students than standardized test scores. In fact, when race and class are ignored in post-secondary GPA predictive models, their effects are often absorbed into the standardized test component, calling into question the validity of such universal standards (Geiser and Santelices, 2007). While critics claim that standardized test scores reflect SES rather than ability, there does seem to be a strong association between test scores and academic potential. One psychology study finds that even when accounting for SES, test scores accounted for 19% of the variation in postsecondary grades (Sackett et al., 2009).
It is unknown whether ACT Math scores are important in college-going because students who achieve in these subjects tend to be more prepared, or whether math scores are important because institutions are better-equipped to serve this group of students. A critical interpretation of results is necessary in order to fully describe college access. Perhaps Math scores are indicative of academic potential, but it may also be the case that institutions are failing to develop talents in other subjects, such as Reading or English.

The vast majority of college admissions counselors are in agreement that more data, rather than less, are desirable in the admissions process. Standardized test scores are often used as an objective measure of success, which can compensate for variations in grade point assessment between schools. Few 4 year colleges do not require the submission of standardized test scores (U.S. News & World Report, 2009). Thus, it is unlikely that many institutions will cease to use any particular measure of educational success altogether. Even so, it may be useful for admissions counselors to know the distinct ways in which ACT scores, GPA, and AP scores can predict success, as well as the biases of each metric.

7.4. URM Access.

They are not fools, they have tasted of the Tree of Life, and they will not cease to think, will not cease attempting to read the riddle of the world. By taking away their best equipped teachers and leaders, by slamming the door of opportunity in the faces of their bolder and brighter minds, will you make them satisfied with their lot? -W.E.B. Du Bois

Racial disparities are often discussed in higher education research. In fact, three of the four policy goals suggested by the ASHE Institutes on Equity and Critical Policy Analysis in 2012 directly addressed racial disparities in higher education (and the study of higher education). Scholars such as Sylvia Hurtado, Mitchell Chang, Kevin Eagan, Jessica Sharkness, and Frances Stage have long studied disparate racial access to higher education.

Studies often focus on individual racial identity rather than the racial makeup of one’s school or district. However, school demographics were more important than individual demographics in all RF prediction models. Interestingly, the RF models and OLS model showed district URM proportion as a positive indicator of postsecondary GPA. However, the effect was not the same for school URM proportion, which RF models show as a negative indicator of academic success. It may be the case that urban districts contain larger URM populations, but largely URM schools within these districts do not necessarily benefit from the district wealth and resources.

7.5. Mobility.

Mobility in the early years of one’s high school career has been shown to have educational benefits for many students (Swanson and Schneider, 1999). No such benefits were found for mobility in the late high school years. Researchers have identified a link
between nonintact families and dropping out, and children from single-parent or step-families are more likely to dropout than others (Astone and McLanahan, 1994). However, there is very little research that includes mobility as a demographic characteristic of interest (Rumberger and Larson, 1998). Through clustering analysis, RF algorithms and linear models, this work identifies mobility as an important predictor of attainment and success in higher education. Over half of the students identified as No-Goers were also mobile. The No-Goers group was also disproportionately Low Income, URM, Limited English and Special Education students. However, mobility is fundamentally different from these identities because one’s mobility may change with time, location, and social context.

From a policy standpoint, mobility may be of particular interest due to it’s relatively high predictive power, lack of previous attention (Rumberger and Larson, 1998), and malleable nature. It is possible that changes in the amount of mobile students in Utah could impact the demographics of educational pathways in the state. It is unclear whether mobility causes students to become No-Goers. However, mobile students are an under-explored phenomenon in Utah, and targeted resources toward this group have the potential to greatly increase college attainment.

Aid to Mobile students must not come at the expense of other marginalized groups. Not only is the No-Goers cluster largely mobile, it also overrepresents URM, Low Income, Limited English and Special Education students. Liberation and equal opportunity cannot be found in the oppression of others; in fact there are large overlaps between many of these marginalized groups. Some of the more frequently studied relationships include that between Special Education status and students of color (Blanchett, 2006), Low Income status and Mobile students (Kerbow, 1996), and Low Income status and students of color (Hu, 2010).


Barriers for Low Income students must be a central topic in the discussion around higher education access. By failing to provide adequate opportunities for educational advancement to Low Income students, the education system is not only doing a disservice to these students, but also losing large numbers of talented students. Previous research has found that high performing Low Income students typically do not have the resources necessary to adequately guide them through the college application process (Kelchen & Goldrick-Rab, 2015). As a result, many Low Income students are unable to attain higher education or achieve highly once enrolled in postsecondary institutions. RF models and linear models suggest that Low Income status, as well as Low Income school and district proportion, are all significant negative predictors of postsecondary attainment and success. However, Pell Eligibility and Pell Grant reception were both significant positive predictors of postsecondary success. The context and implications of each of these findings may be used in order to create pathways for low income students and dismantle barriers to higher education.
Low enrollment numbers from low-income communities may result from uncertainty around the college financing process. Financial information may become available too late in the college application process for many low-income students to realize their academic potential (Kelchen & Goldrick-Rab, 2015). College application behavior varies significantly between high-achieving middle or upper class students and high-achieving low-income students. The majority of low-income students do not apply to a single highly selective institutions, despite the fact that these students may receive more aid at these institutions. Caroline Hoxby and Christopher Avery estimate that while high-achieving, high-income students outnumber high-achieving low-income students 2:1 in the general population, the high-income high achievers outnumber their low-income counterparts 15:1 in college applications to selective institutions. Further, the low-income students who apply to highly selective institutions are equally likely to receive acceptance and progress toward graduation as their more affluent counterparts (Hoxby and Avery, 2012).

In 1976, the original Pell Grant proposal included a bill which would allot awards to institutions which served Pell Grant eligible students. The bill, created under the assumption that educating economically disadvantaged students was socially valuable, but cost more than educating others, would have awarded institutions $2,500 per Pell Grant eligible student. However, the bill was never enacted, and currently there exists no federal financial incentive for universities to educate Pell Grant eligible students (Alexander, 2011). Even so, the number of Pell Grant recipients attending for-profit colleges increased from 1.2 million to 2.1 million from 2008 to 2012, and spending on these students more than doubled during this period, from $3.1 billion to $7.4 billion, in current dollars (Mullin, 2013).

From 2008 to 2012, the Pell Grant program’s expenditures more than doubled, from $14.7 billion to $33.4 billion federal dollars. In this time period, the number of recipients increased 74%, from 5.5 million to 9.7 million. Growth was the result of the larger pool of eligible students, new year-round awards, legislative changes in needs-analysis, and an increased maximum award. Due to the unprecedented growth of the Pell Grant program, legislators have since proposed adjusting the maximum award, time-length of the program, eligible student population, parental wealth and institutions eligible in order to stem or even reduce the program (Mullin, 2013).

Students from low-income backgrounds may benefit disproportionately from early commitment programs, although studies are limited by the small number of these programs in effect (Kelchen and Goldrick-Rab, 2015). The theory behind this finding is that students who are made aware of postsecondary opportunities are more likely to progress in the education system. Similarly, this study shows that students who are made aware of postsecondary Pell Grant funding are more likely to achieve at higher levels of education. Benefits of the Pell Grant award may extend beyond monetary spheres. Financial aid often has the ability to integrate low-income students into the social and academic fabric of campus life (Cabrera, Nora, and Castaneda, 1992).

Methodology around Pell Grant studies includes descriptive statistics (Singell and Stone, 2007), Logit models (Hoxby and Avery, 2012), Probit models (Kelchen and Goldrick-Rab, 2015), Monte-Carlo Simulation (Kelchen and Goldrick-Rab, 2015), Structural Equation Modeling (Cabrera, Nora, and Castaneda, 1992). There is little work
that studies the impacts of Pell Grants using Ensemble Learning methods such as RF. Such studies are important because the RF model is generally used in prediction, a quality which is desirable when estimating the impact of a program or policy. In this study, the RF model concludes that both Pell Grant Reception and Eligibility are positive predictors of postsecondary success (GPA). It is unclear as to why Pell Grant Eligibility would increase a student’s postsecondary success, even if that student did not receive the grant. Future work could examine this phenomenon.

7.7. Critical Contribution.

Following Frances Stage’s model of critical quantitative inquiry, this work makes an effort to (1) further the understanding of current educational inequities using data and (2) challenge the current models being used to assess equity in higher education (2007). In regards to the first objective, this work highlights pathway disparities for Low Income students, Mobile students, and students from largely URM or Low Income schools. While Low Income educational outcomes are often studied, this work reifies the challenges faced by these students, even when performing at the same level of their high school peers. Mobile students are less often discussed in higher education outcome literature, although this identity appears to place students at a significant disadvantage in the pursuit of higher education. The effects of school demographics are potentially more important than the effects of individual identities. In order to understand the disadvantages faced by Low Income and URM students, one can see that discrimination often works at individual, local, and regional levels. The system of college access is complex, but tools such as Random Forest analysis can shed light on real trends and associations. This work challenges models by comparing the predictive power of RF to various linear and logistic models. The RF method is recommended for future critical research, as it presents an interpretable, nuanced understanding of the relationship between variables.

In contrast to previous critical scholarship, which has been largely qualitative, this work places critical theory as necessary in the interpretation of quantitative results. Similarly, this work places quantitative work as a valid approach to critical research. Following Kincheloe and McLaren’s definition of critical theory, this work may be used to describe power relations, examine barriers to autonomy, inspect a variety of identity oppressions, and examine the power of culture and language (2002). Interpretivism is central to critical theory: while the author offers several potential interpretations of research findings, none are to be taken as objective truths. Further interpretations are possible and crucial to a closer understanding of complex systems. Even so, some interpretations align better with the data than others. The critical interpretations provided are supported by the results of this work.

Power associated with race and class is both individual and structural, meaning one may experience advantage or disadvantage based on one’s own identity and the identity of one’s school or district. While cultural power relations are not discussed explicitly in this work, further examination of regional culture could expand current understandings of school and district effects on higher education pathways. Low Income status appears to be a significant barrier in individual and group quests for autonomy and educational
Quantitative inquiry is necessary in advancing understandings of the social world. By combining mathematical and educational theories, this work uses data analysis in order to examine equity of access to higher education and make policy recommendations which may increase equity of access. Perhaps quantifying inequities can bring a society closer to Freire’s liberatory praxis. A relatively underutilized tool in the struggle against oppression, mathematics are crucial in the dismantling of oppressive systems and creation of a democratic society. The author hopes that this work might inspire future collaboration between scholars of education and mathematics. It is imperative that mathematical methods and educational theories exist not only in unison, but harmoniously, as vital organs in the evaluation and progression of humanity.
References.


— “Manual on setting up, using, and understanding random forests v3. 1”. *Statistics Department University of California Berkeley, CA, USA* 1 (2002).


Appendix A. R Code

A.1. Data.

Female Variable Generation:

```R
FEMALE_IND <- rep(0, 43967)
for (i in 1:43967){if (isTRUE(data$ALL_GENDER[i] == "F") & FEMALE_IND[i] == 1)}
```

Institution Type Variable Generation:

```R
attendypub <- rep(0, length(semester))
for (i in 1:length(semester)){if (isTRUE(data$DEGREE_INTENT_CODE[i] == "B") & isTRUE(data$INSTITUTION_TYPE_PUBLICPRIVATE[i] == "Public") & attendypub[i] == 1)}
p <- cbind(p, attendypub)
attendypub <- rep(0, length(semester))
for (i in 1:length(semester)){if (isTRUE(data$DEGREE_INTENT_CODE[i] == "B") & isTRUE(data$INSTITUTION_TYPE_PUBLICPRIVATE[i] == "Private") & attendypub[i] == 1)}
p <- cbind(p, attendypub)
```

Attain College Variable Generation:

```R
AttainCollege <- rep(0, length(data$PERSON_ID))
AttainCollege[match(newd[452]1286741, data$PERSON_ID)] = 1
AttainCollege <- rep(0, length(data$PERSON_ID))
AttainCollege[match(unique(newd[452]1286741, data$PERSON_ID))] = 1
```

Pell Grant Variable Generation:

```R
pell <- rep(0, length(semester))
for (i in 1:length(semester)){if (isTRUE(data$PELL_GRANT_CODE[i] == "A") & isTRUE(data$PELL_GRANT_CODE[i] == "C") & pell[i] == 1)}
p <- cbind(p, pell)
```

Degree Attainment Variable Generation:

```R
grad <- rep(0, length(data$GRADUATION_YEAR))
AttainAssociate <- rep(0, length(data$GRADUATION_YEAR))
AttainBachelor <- rep(0, length(data$GRADUATION_YEAR))
AttainCertificate <- rep(0, length(data$GRADUATION_YEAR))
AttainHigher <- rep(0, length(data$GRADUATION_YEAR))
for (i in 1:length(semester)){if (isTRUE(data$GRADUATION_YEAR[i] == "A") & isTRUE(data$GRADUATION_YEAR[i] == "C") & grad[i] == 1)}
for (i in 1:length(semester)){if (isTRUE(data$GRADUATION_YEAR[i] == "A") & isTRUE(data$GRADUATION_YEAR[i] == "C") & AttainAssociate[i] == 1)}
for (i in 1:length(semester)){if (isTRUE(data$GRADUATION_YEAR[i] == "A") & isTRUE(data$GRADUATION_YEAR[i] == "C") & AttainBachelor[i] == 1)}
for (i in 1:length(semester)){if (isTRUE(data$GRADUATION_YEAR[i] == "A") & isTRUE(data$GRADUATION_YEAR[i] == "C") & AttainCertificate[i] == 1)}
for (i in 1:length(semester)){if (isTRUE(data$GRADUATION_YEAR[i] == "A") & isTRUE(data$GRADUATION_YEAR[i] == "C") & AttainHigher[i] == 1)}
```

STEM Variable Generation:

```R
for (i in 1:length(newd[3])){if (isTRUE(as.numeric(as.character(newd$CURR_CIP_CODE[1])) <= 280000) & isTRUE(as.numeric(as.character(newd$CURR_CIP_CODE[1])) >= 460000)) & STEM[1] != 1)}
for (i in 1:length(newd[3])){if (isTRUE(as.numeric(as.character(newd$CURR_CIP_CODE[1])) <= 420000) & isTRUE(as.numeric(as.character(newd$CURR_CIP_CODE[1])) <= 460000) & STEM[1] != 1)}
for (i in 1:length(newd[3])){if (isTRUE(as.numeric(as.character(newd$CURR_CIP_CODE[1])) <= 420000) & isTRUE(as.numeric(as.character(newd$CURR_CIP_CODE[1])) <= 460000) & STEM[1] != 1)}
for (i in 1:length(newd[3])){if (isTRUE(as.numeric(as.character(newd$CURR_CIP_CODE[1])) <= 420000) & isTRUE(as.numeric(as.character(newd$CURR_CIP_CODE[1])) <= 460000) & STEM[1] != 1)}
```

Computing District and School level Metrics:

```R
school <- rep(0, length(data$SCHOOL_ID))
unmeanschool <- aggregate(x = data$SCHOOL_ID, by = list(data$SCHOOL_ID), mean)
hist(unmeanschool$x)
summary(unmeanschool)
```
Merging Data:
```
data-merge(data[,c(1,1,6,27,19,38,89,99,47,50,71,72,43)],.new4[,c(1,1,16,175)],by="PERSON_ID",all=TRUE)
data-merge(data[,c(1,12,18,41,200,206)],.new4[,c(1,13,174)],by="PERSON_ID",all=TRUE)
length(person)
data[1:10,11:17]
```

GPA for Missing and Non-Missing ACT Scores:
```
### show that two groups are different
summary(data[which(data$ACTComposite == 0),]
summary(data[which(data$ACTComposite == 0),]
```

Amelia Missing Variable Imputation:
```
### Amelia ###
`Amelia`
data.out-amelia(data$complete.cases(data$SUMMARY_GPA),c(1,1,7,11,30,31,33,34,282:286,208:211)),bounds = matrix(c(11,0,12,0,11,0,16,6)),
```

A.2. Dimensionality Reduction and Data Exploration.
```
Data Restructuring by Dates:
p-`order`([PENTRY_DATE,],)
newdates-as.Date(PENTRY_DATE,format="Wn/Wn/WY")
as.numeric(as.Date("2020-12-31"))
new3-`cbind`(new2,newdatesnumeric)
new3`numeric`[as.numeric(new3) == NULL]
newdatesnumeric-as.numeric(newdates)
as.numeric(as.Date("1900-3-2")) == 0
`cut`
`seq.Date`
cuts-`cut`[as.Date("1999-7-1"),as.Date("2013-12-1"),by="4 months"]
cuts[28]
cut(newdatesnumeric,cuts)
cut,dates-out.Date(newdates,breaks=cuts)
new3-`cbind`(new3,out,dates)
```

Enrollment Type:
Re-coding Data to Reflect All Possible Semesters:

```r
fulltime<-as.numeric(pENROLLMENT_STATUS_CODE)
summary(pENROLLMENT_STATUS_CODE)
for(i in 1:length(p[i,1])){
  if(isTRUE(as.character(pENROLLMENT_STATUS_CODE[i])=="F")){fulltime[i]=4}
  if(isTRUE(as.character(pENROLLMENT_STATUS_CODE[i])=="R")){fulltime[i]=0}
  if(isTRUE(as.character(pENROLLMENT_STATUS_CODE[i])=="W")){fulltime[i]=2}
  if(isTRUE(as.character(pENROLLMENT_STATUS_CODE[i])=="L")){fulltime[i]=1}
  if(isTRUE(as.character(pENROLLMENT_STATUS_CODE[i])=="")){fulltime[i]=3}
}
PCA Analysis with Re-coded Data:

```}

A.3. Clustering Pathways.

Flexclust K-medoids Clustering:

```r
## Flexclust Clustering ##
"?flexclust-class"
#mcooky$SMMX<cca(new7,-1)[complete.cases(data2$CUMULATIVE_GPA),],k=21, family= kccaFamily("kmedlans"), save.data = T)
summary(mcooky$SMMX[1])
clusters(mcooky$SMMX)
info$mcooky$SMMX("size")
info$mcooky$SMMX("av_dist")
```

Clustering According to STEM & GPA:
A.4. Predicting Pathways and Success.

Pathways Random Forest and Partial Plots:

```r
RFa=randomForest(x~v, data=complete casing, formula=cluster ~ complete, ntree=100, mtry=3, importance=TRUE)
predictions=RFa$predict(Complete)
plot(RFa$importance, main='Variable Importance for Pathway Classification', ylab='

College Attainment Random Forest:

```r
RFa=randomForest(x~v, data=complete casing, formula=cluster ~ complete, ntree=100, mtry=3, importance=TRUE)
predictions=RFa$predict(Complete)
plot(RFa$importance, main='Variable Importance for College Attainment Classification', ylab='

Creating Test and Training Sets for GPA Estimation:

```r
set.seed(123)
data=sampleTrain(

A.5. Model Comparisons for Predicted Student Success.

```
Linear Models Generation:

### Predicting GPA only college student:
```
goal = glm(y ~ female + all.income + cumulative.gpa + low.income.ind + sped.ind + mobi.ind + migrant.ind + limited.english.ind + attend.college + pat1 + p1 + p2 + pgra
summary(goal)
predictions = predict(gpa,newdata = data2[match(newPERSON_ID, data2$PERSON_ID), c(5,7,12,24,26,28,30,37,41,200:2005)]
summary(predictions)
```

Logarithmic Model Generation:
```
log = glm(log(gpa ~ female + all.income + cumulative.gpa + low.income.ind + sped.ind + mobi.ind + migrant.ind + limited.english.ind + attend.college + pat1 + p1 + p2 + pgra
summary(log)
predictions = predict(log,newdata = data2$PERSON_ID,c(5,7,12,24,26,28,30,37,41,200:2005)]
summary(predictions)
```