

A TOOL FOR WHO WILL DROP OUT OF SCHOOL

By

Colette Joelle Barca, Bachelor of Science Degree in Mathematics

A thesis submitted to the Graduate Committee of
Ramapo College of New Jersey in partial fulfillment
of the requirements for the degree of
Master of Science in Data Science
Spring, 2022

Committee Members:

Osei Tweneboah Ph.D., Advisor

Amanda Beecher Ph.D., Reader

Debbie Yuster Ph.D., Reader

COPYRIGHT

© Colette Joelle Barca

2022

Dedication

To my parents, Gregory and JoAnn.

Thank you for encouraging me to pursue my dreams and to find a career that excites and inspires me. I would not be here today without your constant support and encouragement.

To my brother, Thomas.

Thank you for sacrificing our time together so I could keep up with my work and stay focused.
Thank you for believing in me.

To students struggling in school.

You are seen. You are heard. You are important.

Acknowledgements

First and foremost, I would like to thank my advisor, Professor Osei Tweneboah, for his constant guidance, patience, and support throughout this entire process. Thank you for taking the time to meet with me each week and for helping me break down this project into small, digestible pieces. I am so grateful I was able to work with and learn from you during my final semester as a Ramapo College student.

I would also like to thank the other members of my thesis committee, Professors Amanda Beecher and Debbie Yuster. Thank you for helping me anticipate various issues I might (and did) encounter along the way and for providing numerous suggestions as to how I could address those problems.

I would like to express my deepest gratitude to the Ramapo College Upward Bound Math Science Program Director, Dr. Sandra Suarez, without whom this research would be impossible. Because of her immense generosity and meticulous data maintenance, we had access to quality data that allowed us to meaningfully analyze college dropout rates.

I would like to extend a special thank you to the entire Ramapo College Data Science faculty for providing me the tools, skills, and resources needed to conduct this research. Thank you for all your support and for giving me so many opportunities to grow as a data scientist.

Lastly, I would like to thank Ms. Allison Carluccio for helping me prioritize my tasks and preventing me from getting overwhelmed.

Table of Contents

DEDICATION	iv
ACKNOWLEDGEMENTS	v
TABLE OF CONTENTS	vi
LIST OF TABLES	ix
LIST OF FIGURES	xii
LIST OF APPENDICES	xiii
ABSTRACT	1
CHAPTER	
I. INTRODUCTION	2
II. BACKGROUND	5
III. METHODOLOGY	8
About the Datasets	8
Upward Bound	8
High School Information	10
National Center for Education Statistics	10
New Jersey Department of Education	11
Federal Student Aid	11
Esri	12
Data Cleaning and Preparation	13
Research Design	21

Distance-Based Techniques	22
Clustering-Based Techniques	23
Classification-Based Techniques	23
Assumptions	24
Variable Selection	25
Exploratory Data Analysis	25
Modeling	28
Modeling with Distance-Based Techniques	28
<i>K</i> -Nearest Neighbors (KNN)	28
Pruning	30
Modeling with Clustering-Based Techniques	31
<i>K</i> -Means	31
Agglomerative/Hierarchical	32
Gaussian Mixture Models (GMM)	34
Modeling with Classification-Based Techniques	36
Stochastic Gradient Descent (SGD) and Logistic Regression	36
Support Vector Machines (SVM)	38
Decision Trees, Random Forests, and Boosting	40
Voting Classifiers	43
IV. ANALYSIS AND DISCUSSION	46
Analysis	46
Discussion	65

V. CONCLUSION	73
Contributions	74
Future Work	76
Density-Based Anomaly Detection	77
Assessing for Differences in Graduation Rates Across Several Years	77
Reassess Students at the Culmination of Each School Year	77
Grid Search	78
Combining the Supervised and Unsupervised Models	78
Determining When to Raise a Warning for a Potential Dropout	79
REFERENCES	81
APPENDICES	87

List of Tables

1. Academic Variables and Their Possible Values (If Categorical)	17
2. Institutional Variables and Their Possible Values (If Categorical)	18
3. Personal Information Variables and Their Possible Values (If Categorical)	19
4. Social Demographic Variables and Their Possible Values (If Categorical)	20
5. Number of Outliers Identified by Various k -Nearest Neighbors Models	29
6. Best Performance Attained by Each Distance-Based Model	30
7. Number of Dropouts Predicted by Various k -Means Models	32
8. Number of Dropouts Predicted by Various Agglomerative/Hierarchical Models	34
9. Number of Dropouts Predicted by the Gaussian Mixture Models	35
10. Best Performance Attained by Each Clustering-Based Model	36
11. Number of Dropouts Predicted by Various Stochastic Gradient Descent and Logistic Regression Models	38
12. Number of Dropouts Predicted by Various Support Vector Machine Models	39
13. Best and Worst Performance Attained by Each Tree-Based Technique	43
14. Number of Dropouts Predicted by Each Voting Classifier Model	44
15. Best Performance Attained by Each Classification-Based Model	45
16. Best Distance-Based Technique's Performance as an Anomaly Detection Tool	47
17. Best Clustering-Based Technique's Performance as an Unsupervised Anomaly Detection Tool	47
18. Best Distance-Based Technique's Characteristics of Students Most Likely to Drop Out	49
19. Best Clustering-Based Technique's Unsupervised Characteristics of Students Most Likely to Drop Out	50

20.	Best Distance-Based Technique's Characteristics of High Schools Most Likely to Create Dropouts	52
21.	Best Clustering-Based Technique's Unsupervised Characteristics of High Schools Most Likely to Create Dropouts	53
22.	Best Clustering-Based Technique's Unsupervised Characteristics of Students Most Likely to Graduate	54
23.	Best Distance-Based Technique's Characteristics of Students Most Likely to Graduate	54
24.	Best Distance-Based Technique's Characteristics of High Schools Most Likely to Create Graduates	55
25.	Best Clustering-Based Technique's Unsupervised Characteristics of High Schools Most Likely to Create Graduates	55
26.	Best Clustering-Based Technique's Performance as a Supervised Anomaly Detection Tool	56
27.	Best Classification-Based Technique's Performance as an Anomaly Detection Tool	57
28.	Best Clustering-Based Technique's Supervised Characteristics of Students Most Likely to Drop Out	58
29.	Best Clustering-Based Technique's Supervised Characteristics of Students Most Likely to Graduate	59
30.	Best Classification-Based Technique's Student Characteristics Most Associated With Graduation Status	60
31.	Best Classification-Based Technique's High School Characteristics Most Associated With Graduation Status	61
32.	Best Clustering-Based Technique's Supervised Characteristics of High Schools Most Likely to Create Graduates	62
33.	Best Clustering-Based Technique's Supervised Characteristics of High Schools Most Likely to Create Dropouts	63
34.	Performances of the Final Models on the Testing Set	64
35.	The Twenty Most Important Features of Each Final Model (In Order of Importance)	68

36.	Performance of the Terminal Model on the Model Deployment Set	69
37.	A Comprehensive List of Which Factors Each Technique Deems Most Associated With Graduation Status	71

List of Figures

1. Example of a Skewed Distribution	14
2. Example of a Normal Distribution	15
3. Dataset's Graduation Distribution	26
4. Dataset's Gender Distribution	26
5. Distribution of Graduation by Gender	26
6. Level of Crime in Each High School ZIP Code (9 ZIP Codes in Total)	27
7. Example of a Dendrogram	32
8. Example of a Decision Tree	41
9. Confusion Matrix for Column Set A's Boosting Model	57
10. Confusion Matrix for Final Model #1: Testing Set	65
11. Confusion Matrix for Final Model #2: Testing Set	65
12. Confusion Matrix for Final Model #3: Testing Set	65
13. The Ten Most Important Features in Model #1	66
14. Confusion Matrix for the Best Model on Model Deployment	69

List of Appendices

A. Variables in the Final Dataset	87
B. Using Linear Algebra to Calculate the Geographic Distances	90
C. Description of Each Variable	91
D. Variables in Column Set A	100
E. Variables in Column Set B	103
F. Variables in Column Set C	106
G. Outcome of Each k -Nearest Neighbors Attempt	109
H. Outcome of Each Pruning Attempt	110
I. Outcome of Each k -Means Attempt	111
J. Outcome of Each Agglomerative Attempt	112
K. Outcome of Each Logistic Regression Attempt	114
L. Outcome of Each Stochastic Gradient Descent Attempt	115
M. Outcome of Each Support Vector Machine Attempt	117
N. Outcome of Each Decision Tree Attempt	120
O. Outcome of Each Random Forest Attempt	121
P. Outcome of Each Boosting Attempt	122
Q. Outcome of Each Voting Classifier Attempt	124

Abstract

A student's high school experience often forms the foundation of his or her postsecondary career. As the competitive nature of our nation's job market continues to increase, most businesses stipulate all applicants need a college degree. However, recent studies show approximately one-third of the United States' college students never obtain a degree. Although colleges have developed methods for identifying and supporting their struggling students, early intervention could be a more effective approach for combating postsecondary dropout rates. This project seeks to use anomaly detection techniques to create a holistic early detection tool that indicates which high school students are most at risk to drop out of college. An individual's high school experience is not confined to the academic components. As such, an effective model should incorporate both environmental and educational factors, including various descriptive data on the student's home area, the school's area, and the school's overall structure and performance. This project combined this information with data on students throughout their secondary educational careers (i.e., from ninth through twelfth grade) in an attempt to develop a model that could detect during high school which students have a higher probability of dropping out of college. The clustering-based and classification-based anomaly detection algorithms detail the situational and numeric circumstances, respectively, that most frequently result in a student dropping out of college. High school administrators could implement these models at the culmination of each school year to identify which students are most at risk for dropping out in college. Then, administrators could provide additional support to those students during the following school year to decrease that risk. College administrators could also follow this same process to minimize dropout rates.

Chapter One: Introduction

More and more, recruiters and businesses are requiring all applicants to have a college degree. According to Forbes magazine, “over the past decade, [businesses] have begun to demand a bachelor’s degree in hiring workers for jobs [including those] that traditionally [have not] required one” (Fuller, 2017). And yet, the college dropout rates are increasing. In fact these rates are rising so dramatically that “[in] American higher education...about one in three students who enroll in college never earn a degree” (Leonhardt & Chinoy, 2019). This has sparked a chain reaction of responses from numerous groups of individuals, beginning with parents. According to the New York Times, parents and guardians are advising their children to avoid applying to colleges with lower graduation rates (2019). Naturally, colleges want to avoid being labeled as one of these “low-rated” schools; thus, they have developed methods for identifying and supporting students. These institutions are hiring individuals to help them develop and implement an intervention process. Currently, these methods are heavily-based around academic performance.

The issue of increasing dropout rates has become so prevalent that data scientists and psychologists also feel compelled to look into this situation to try and discover what is going on. We will discuss the specific results and findings of their research in the next chapter. Briefly, numerous studies conducted over the last twenty years continue to report similar findings; the education system’s current methods for reducing dropout rates do not work. When we only identify at-risk students after they display poor academic performance, it becomes very difficult to help students recover from that deficit. Because we rely on such a narrow set of indicators, we ultimately provide the support services too late in the student’s academic career.

In an effort to provide a new approach to this problem, we want to see if it is possible to use anomaly detection techniques to create a holistic early detection tool that indicates which high school students are most at risk to drop out of college. Within this project, we seek to develop an anomaly detection algorithm that details the situational and academic circumstances that most frequently result in a student dropping out of college. If this project results in an effective early detection tool, high school administrators could implement the model at the culmination of each school year to identify which students are most at-risk for dropping out in college. Then, administrators could provide additional support to those students during the following school year to decrease that risk. Even if this model cannot achieve early detection efficacy, college administrators could still follow this same process to minimize dropout rates. In either scenario, this model would benefit both institutions and students.

To effectively address this problem, we need data that details a student's personal, demographic, and academic information while the individual attended high school along with information regarding what happened to those same students when they attended college. The Ramapo College Upward Bound Math Science program's Annual Performance Reports contain both of these components. In Chapter Three, we will provide more detail on these reports. For context, Upward Bound is part of the Federal TRIO Programs. As the Upward Bound website explains:

Upward Bound provides fundamental support to participants in their preparation for college entrance. The program provides opportunities for participants to succeed in their precollege performance and ultimately in their higher education pursuits. Upward Bound serves: high school students from low-income families; and high school students from families in which neither parent holds a bachelor's degree. The goal of Upward Bound is

to increase the rate at which participants complete secondary education and enroll in and graduate from institutions of postsecondary education. (*Upward Bound: Purpose*, 2021)

Evidently, Upward Bound is designed for a very specific student demographic population. In fact, TRIO even stipulates specific proportions program directors must adhere to when admitting students. Each year, “two-thirds of the participants in a project must be both low-income and potential first-generation students. The remaining one-third must be either low-income, first-generation college students, or students who have a high risk for academic failure” (*Upward Bound: Purpose*, 2021). Upward Bound students are the perfect demographic to analyze for this research, as the dropout rates of children from low-income families “(9.4 percent), [are] higher than the rates of their peers from middle-income (5.4 percent) and high-income (2.6 percent) families” (McFarland, Cui, & Stark, 2018, p. 11). Of course, the model could be implemented in *any* district, as dropping out is not an issue that is unique to a particular population.

Based on our objective, the layout of this thesis is as follows: Chapter One gives a preview of the thesis topic under consideration and a brief introduction to the Upward Bound program. Chapter Two contains the review of related literature. In this chapter, we shall give a brief overview of various qualitative and quantitative research pertaining to college dropout rates that were conducted within the past twenty years. In Chapter Three, we discuss several anomaly detection techniques and outline the various datasets used to conduct this research. Detailed explanations of our numerous modeling attempts are also included in this chapter. In Chapter Four, we focus on the analysis of data and discussion of findings. Conclusions are contained in Chapter Five.

Chapter Two: Background

In Chapter One, we alluded to various psychological and analytical studies concerning college dropout rates. In this chapter, we will delve into their research and highlight some of their important findings. This will provide context for our primary research objective and will supply a clear understanding of how this project expands upon the work of these researchers.

In 2000, a pair of psychologists decided to conduct a qualitative study regarding college dropouts and found “that students who drop out of school suffer from a host of negative consequences, ranging from high unemployment and low earnings to poor health and increased criminal activity” (Rumberger & Thomas, 2000). They also discovered “that students who are not engaged in school and do not attend regularly are more likely to have low test scores and a higher risk of dropping out” (Rumberger & Thomas). This suggests that low academic performance and dropping out of school are side effects, rather than the source of an issue. Building upon this, they uncovered it is “the characteristics of schools, as well as the characteristics of students [that] influence...the final outcomes of test scores and dropout rates” (Rumberger & Thomas).

More recently, a second group of psychologists performed a different qualitative study to assess the causes of dropping out from an ecological perspective. The researchers “found the factors related to school dropout included adjustment problems at the individual level; family, teachers, and economy at the microsystem level; and gender at the macrosystem level” (Muharrem, et al., 2020). That same year, a group of data scientists organized a quantitative study in which they used linear regression to predict a student’s level of risk for dropping out. They too found that dropping out “is not solely due to the characteristics of the students, but that

there are other factors such as the school environment that lead to dropout” (Zorbaz & Özer, 2020). Even twenty years later, researchers are reporting essentially the same findings: it is economic and environmental factors, rather than academic ones, that are most associated with dropping out of college.

However, current identification methods rely on academic performance metrics, such as standardized test scores and grade point averages (GPA). Although test scores are among the most common indicators of success, they are insufficient tools for determining a student’s academic abilities (or lack thereof) (Hiller, 2019). While these exams are often portrayed as assessments of student knowledge, their structure and design inadvertently make them assessments of strategy and test taking skills. They are “essentially designed to measure how well one is able to take a standardized test, while also being under the pressure of time;” as such, “essentially designed to measure how well one is able to take a standardized test, while also being under the pressure of time” (Hiller). Furthermore:

Any assessment, standardized or not, is not some magical fortune-telling device. If properly designed and tested, an assessment can partially and probabilistically predict a particular outcome. The wider the range of desired outcomes, the less relevant any particular test will, or should, be. Of course, the particular type of academic preparation that SATs and ACTs measure can represent a reasonable component of what should matter to academic institutions as they select students. But as predictors of the capacity to conduct...academic study...these tests assess only one narrow part of applicants’ capacities. (Shulman, 2018, p. 9)

The previous research emphasizes dropping out of school is correlated with numerous non-academic factors. Rather than only using academic performance to identify at-risk students, detection tools should also consider environmental, economic, and demographic information.

In this chapter, we highlighted the major takeaways from years of research regarding college dropout rates. Primarily, we found that throughout the last twenty years, researchers from various fields identified environmental, economic, and demographic information as most strongly associated with dropping out. This indicates that the current academically-focused identification methods are insufficient and should be augmented with holistic approaches. In the next chapter, we will reveal how we incorporated this information into our anomaly detection models.

Chapter Three: Methodology

In Chapters One and Two, we presented the primary goal of this thesis and placed our research into the context of current approaches to and understandings of college dropout rates. In this chapter, we will detail how we used anomaly detection techniques in an attempt to create an early detection tool that indicates which high school students are most likely to drop out of college. We will provide a more detailed explanation of the Upward Bound data briefly mentioned in Chapter One. We will also describe the various datasets we incorporated to directly address the psychological findings discussed in Chapter Two, along with how we combined all the information into one comprehensive dataset. We will outline how we designed our research, including any assumptions made along the way. Finally, we will present the findings of our exploratory data analysis and modeling efforts.

About the Datasets

In this section, we will describe the various datasets used for this research, detail why we included them in our project, and identify the information we extracted from each source.

Upward Bound

The student-specific data for this research came from the Ramapo College Upward Bound Math Science program. Specifically, we utilized their historic Annual Performance Reports. Henceforth, we shall refer to this organization as simply Upward Bound. In Chapter One, we described the purpose of the program and provided a brief overview of its yearly activities. As we mentioned, a major focus of Upward Bound is its dedication toward college preparation and helping students “graduate from institutions of postsecondary education”

(*Upward Bound: Purpose*, 2021). To assess the program’s ability to meet this goal, TRIO¹ requires Upward Bound directors to submit an Annual Performance Report each November (*Upward Bound: Performance*, 2021). Within these reports, directors must indicate how many of their students “[attain] either an associate or bachelor’s degree within six years following graduation from high school” (Office of Postsecondary Education, 2021). Because these reports are submitted on an annual basis, each report documents students from ten distinct cohorts, where a cohort corresponds to a single graduation class. For example, the 2022 Annual Performance Reports will include students from the 2016 through 2025 secondary graduating classes.

Although many Upward Bound students graduate in fewer than six years, it was important we only analyzed cohorts that completed the six-year period prior to the start of this project (i.e., prior to January 2022). This way, we avoided penalizing students who took longer than four years to graduate, inadvertently skewing our results. Thus, the students who graduated high school in 2015 encompassed the most recent cohort included in this research. Additionally, Ramapo’s Upward Bound program first began digitizing their reports in November 2008. Our modeling techniques required digitized versions of the datasets; hence, seniors who graduated at the end of the 2007–2008 school year constituted the earliest graduating class within our dataset. Still, numerous portions of the Upward Bound data came from several boxes of paper data manually entered prior to modeling.

In total, we used fourteen Annual Performance Report files to create our dataset (Ramapo College Upward Bound Math Science Program, 2008–2021). This consisted of any students who graduated high school between 2008 and 2015; these students were in the 2014 to 2021

¹ The Federal TRIO Programs (TRIO) are a collection of “eight programs [including Upward Bound] targeted to serve and assist low-income individuals, first-generation college students, and individuals with disabilities” (*Federal TRIO*, 2022). Please refer to the TRIO webpage for further information.

postsecondary cohorts, respectively. There were one-hundred forty-six students within our final dataset. These reports contained demographic, academic (both secondary and postsecondary), economic, and Upward Bound-specific information for each student. Although these reports included eighty-three fields (nine of which we manually collected and entered), we utilized fifty-two of these features in some manner throughout this project, keeping only twenty-seven in the final dataset to use for modeling. These twenty-seven variables are highlighted in red within Appendix A.

High School Information

In order to conduct a comprehensive assessment of a student's high school experience, it was important to incorporate information pertaining to its most prominent aspect: the classroom environment. As such, we wanted to include data that reported each school's general structure and performance so we could analyze a high school's effect on college graduation status.

National Center for Education Statistics. According to their website, the "National Center for Education Statistics (NCES) is the primary federal entity for collecting and analyzing data related to education in the U.S." (NCES, n.d.). Their public datasets include information pertaining to all schools within the United States and can be filtered for specific schools or entire districts. Their reports provide the complete address of each institution and briefly describe the school's population and academic design. We used the NCES records for the Paterson and Newark Public Schools Districts (2021c, 2021a) along with that of the Passaic County Technical Institute Public School (2021b) to collect this information for all the high schools within our dataset. We used seventeen of their twenty-three fields in some capacity during this project and directly used seven for modeling. These seven predictors are highlighted in orange within Appendix A.

New Jersey Department of Education. The New Jersey Department of Education (NJDOE) creates and publishes annual performance reports for each school within the state “to help evaluate whether all students have equitable access to high quality education” (NJDOE, n.d.). They condense all these reports into a single publicly available file. Schools are assessed using a variety of metrics including: enrollment trends, student performance on standardized tests, course offerings, and student participation. The file contains a total of forty-six reports, though we only used a fraction of this information. We focused on these reports’ social demographic information indicating the percentages of each school’s population that belong to various demographic groups (NJDOE, 2020). Of the fourteen fields initially extracted from these files, we used thirteen for modeling. These thirteen predictors are highlighted in yellow within Appendix A.

Federal Student Aid

Because this project centers around the potential creation of an early detection tool, we avoided incorporating descriptive statistics of students’ colleges into our dataset. We acknowledge that a student’s college environment can also influence a student’s decision to graduate or drop out of college; however, this is outside the scope of this research. Instead, we used the Federal Student Aid’s public file containing a comprehensive list of the federal school codes (which corresponded to the college codes included in the Upward Bound data) to retrieve the addresses of all the colleges within our dataset (2022). We used these addresses to compute an additional metric, which will be discussed later in this chapter. We utilized six of the file’s nine fields at least once throughout this project, but we removed all of these prior to modeling.

Esri

To diversify the environmental components of our dataset, we needed to incorporate descriptive data on both the student's home and high school's areas. In Chapter Two, we mentioned a study that indicated many dropouts encounter "increased criminal activity" (Rumberger & Thomas, 2000, p. 40). As such, we felt obtaining and incorporating quantitative data pertaining to crime within these two locations would be the most effective way to integrate this finding into our project. Esri's public "ZIP Code" dataset indicates an index level associated with various types of crime for every ZIP Code in the United States. The dataset "incorporates information from the [Applied Geographic Solutions] national CrimeRisk database" (Esri, 2019). The database is "based on an extensive analysis of several years of crime incidents reported by most [U.S.] law enforcement jurisdictions" (Esri). From the description of the dataset's corresponding map, we find the indexes are defined such that:

The index values for the [U.S.] level are 100, representing average crime for the country. A value of more than 100 represents higher crime than the national average, and a value of less than 100 represents lower crime than the national average. For example, an index of 120 implies that crime in the area is 20 percent higher than the [U.S.] average; an index of 80 implies that crime is 20 percent lower than the [U.S.] average. (*Crime in the United States*, 2021)

For clarity, simply subtracting each index value by one hundred would equate to the area's percentage of crime with respect to the average. Continuing with the same examples from above, if we subtract one hundred from one-hundred twenty, we get positive twenty. Thus, the area's crime rates are twenty percent above average. Similarly, subtracting one hundred from eighty yields negative twenty; the crime is twenty percent below average.

The dataset contains twenty-nine attributes, including index and aggregate values for ten crime categories Esri refers to as “serious” (2019). After filtering the dataset to only include New Jersey ZIP Codes, we extracted thirteen of these fields, excluding all the aggregate features. We used twelve of these thirteen fields for modeling. These dozen appeared twice in our final dataset: once for the high school’s area and again for the student’s area. The twenty-four predictors are highlighted in green within Appendix A. We also used the ten index attributes to create categorical variables. We categorized any index values greater than one hundred as “High Crime Areas”, those less than one hundred as “Low Crime Areas”, and any equal to one hundred as “Average Crime Areas”. By incorporating both quantitative and qualitative metrics for crime, we could assess if dropout rates corresponded with larger index values or if merely being in an area with above-average crime was sufficient (and vice versa).

Data Cleaning and Preparation

We spent a significant amount of time collecting and preparing the data for modeling. This involved combining the five datasets, addressing missing data, and transforming portions of the dataset, as needed. By merging the datasets, we condensed all their information into one comprehensive table with one-hundred forty-six rows and one-hundred seven columns. Each column contained the information of a single variable for every student in the dataset; each row described a single student using all the variables.

As we mentioned in the previous section, we did not directly use all fields pulled from the datasets for modeling. Instead, we used many of these features to create other variables. For example, we wanted to assess whether the distance between a student’s home and school (both high school and postsecondary) influenced graduation status. These three addresses came from the Upward Bound, NCES, and Federal Student Aid data, respectively. Using the Google Maps

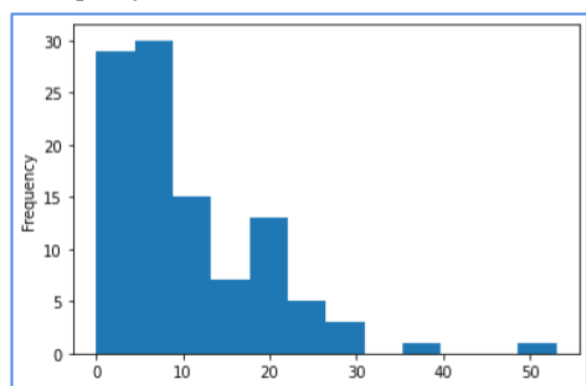
API, we converted these addresses into geographic coordinates. This information contained the corresponding latitude and longitude values, which we used to compute each location's three-dimensional coordinates. Finally, we used linear algebra to calculate the distance between each student's home and schools in kilometers. A more in-depth discussion of this process is provided in Appendix B. We only kept the two distances in the final dataset, even though we used approximately twenty variables to compute them.

Once we completed all our computations, we inspected the dataset for any missing values. In total, there were eight categorical and four continuous variables missing data. Categorical “variables take on values in one of K different classes, or categories” (James, et al., 2013, p. 28), indicating which group the instance belongs to. For example, our `Grade_EnteredUB` variable tells us if the student first attended Upward Bound as a freshman, sophomore, junior, or senior. These variables are commonly referred to as qualitative, as they describe a specific quality. We replaced missing categorical values with the variable's most frequent category (i.e., the mode).

Continuous “variables take on numerical values” (James, et al., 2013, p. 28) and typically express an amount. The `Absences` variable denotes the number of times a student missed a day of school during his or her senior year of high school. These variables are commonly referred to as quantitative, as they provide a quantity. Unlike categorical variables, there is no universal method for replacing missing continuous values; the process is dependent on the shape of the dataset. When the majority of

Figure 1

Example of a Skewed Distribution



the data is clumped around the high-end or low-end of the range of values, like the data in Figure 1, the distribution is skewed. In these instances, we imputed missing values with the variable's median, as it is more representative of the dataset's typical values. When the majority of the data is clumped in the middle of the range of values, as in Figure 2, the data is normally distributed.

In these situations, we replaced missing values with the mean, as this metric is an accurate depiction of the variable's "average" value.

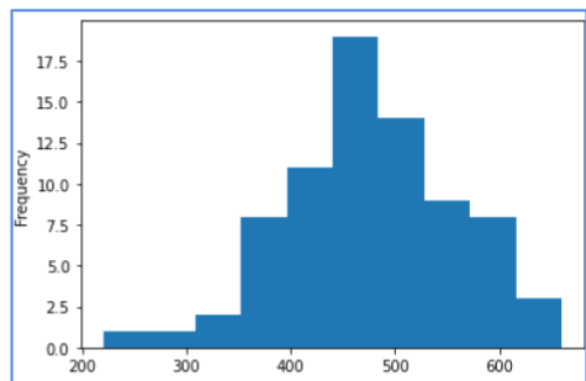
Before modeling, we had to reformat portions of our dataset. Typically, "Machine Learning algorithms prefer to work with numbers" (Géron, 2019, p. 66), rather than with

words. To accommodate this, we converted all textual data to numbers using ordinal and one-hot encoding. If the possible values of a qualitative variable have a clear ranking or order, we use ordinal encoding to reflect this hierarchy. There were twenty such variables within our dataset, the twenty categorical features we created using the crime data. The `OrdinalEncoder` class from Scikit-Learn informed our models that there is a relationship between each variable's values, a notion that we as individuals are able to interpret textually.

For the four remaining text-based variables, we employed Scikit-Learn's `OneHotEncoder` class, which creates "one binary attribute per category" (Géron, 2019, p. 67). This means one feature is made for each distinct value within a column. For example, our dataset's `HS_ZIP` column contained nine different ZIP Codes, corresponding to the locations of each high school. The encoder used these values to create nine dummy variables. These variables only have two possible values: one and zero. These values translate to "yes" and "no",

Figure 2

Example of a Normal Distribution



respectively. In each row, “only one attribute will be equal to 1 (hot), while the others will be 0 (cold)” (Géron, p. 67). So, all students who attended Barringer High School in Newark had a one in the `HS_07104` column and a zero in the remaining eight columns. After completing all these steps, our final dataset contained one-hundred forty-six rows, each corresponding to a student described by one-hundred seventy-three columns. The table in Appendix C provides a brief description of each variable.

By reviewing the information provided by each column, we can categorize all one-hundred seventy-three variables into four distinct categories. Doing so will help us form meaningful interpretations of our models’ results, allowing us to identify which categories impact a student’s college graduation status. Each variable provides either academic, personal, institutional, or social demographic information. These segmentations illustrate how this project aims to analyze influences on college dropout rates from all possible angles. The entries within the following four tables summarize the possible values of all the predictors. Unlike the table in Appendix A, these tables do not include our dataset’s exact variable names. Instead, they provide a brief description of the information gained from each feature or its possible values. Each table is divided into three or four subcategories to increase their interpretability and highlight the nuances within each information category. Table 1 displays the variables concerning an individual’s academic experience organized by the student’s needs, abilities in relation to the state’s standards, academic performance, and academic involvement. Table 2 includes all the variables associated with the dataset’s schools. These features outline basic information on every school, their students, and the areas they are located in. Since these variables are not dependent on the dataset’s students, their values will be identical for every student who attended the school. Table 3 captures all the variables associated with students’ lives outside of school, including their

extracurriculars, home neighborhoods, and how far away they live from high school and college. The final table, Table 4, contains only a small number of features, but its information could help us apply our findings to specific populations, rather than restricting us to only specific students.

Table 1

Academic Variables and Their Possible Values (If Categorical)

Academic Information			
Issues & Needs	Proficiency	Quantitative Metrics	Participation
Diagnosed With a Learning Disability	Graduated From College	Grade Point Average (GPA)	Active in Upward Bound All Year
From a Predominantly Low Income Community	Met the State's Literary Proficiency Standards	SAT Math Score	Active in Upward Bound During the School Year
Interested in Math/Science	Met the State's Math Proficiency Standards	SAT Reading Score	Active in Upward Bound During the School Year & Summer Bridge
Lack of Career Goals	Obtained a Bachelor's Degree	SAT Writing Score	Active in Upward Bound During the Summer Bridge
Lack of Confidence/Self Esteem/Social Skills	Obtained a STEM Degree		Active in Upward Bound During the Summer Component
Lack of Opportunity/Support	Obtained an Associate's Degree		Grade Level Upon Entry to Upward Bound
Limited English Proficiency	Took at Least One Honors-Level Course		Inactive in Upward Bound During Senior Year
Low Educational Aspirations	Took at Least One Advanced Placement (AP) Course		Number of Absences
Low GPA			
Low GPA & Low Educational Aspirations			
Low GPA & Low Test Scores			
Low Test Scores			
Low Test Scores & Low Educational Aspirations			
Pre-Algebra or Algebra Not Completed by 10th Grade			

Table 2*Institutional Variables and Their Possible Values (If Categorical)*

Institutional Information		
Institutions & Details	High School Population's Social Demographics	High School's Area
College Attended	Percentage of Females	Area's Population
High School Attended	Percentage of Males	Size of the Area (ZIP Code)
High School's Student-Teacher Ratio	Percentage of Economically Disadvantaged Students	Assault Index & Rates
High School Is a Title I School	Percentage of Students with Disabilities	Burglary Index & Rates
High School Is Title I School-Wide	Percentage of English Language Learners	Larceny Index & Rates
Number of Free Lunch Students	Percentage of Homeless Students	Murder Index & Rates
Number of Reduced Lunch Students	Percentage of Students in Foster Care	Personal Crime Index & Rates
Number of Students at High School	Percentage of White Students	Property Crime Index & Rates
Number of Teachers at High School	Percentage of Hispanic Students	Rape Index & Rates
	Percentage of African American/Black Students	Robbery Index & Rates
	Percentage of Asian Students	Total Crime Index & Rates
	Percentage of Hawaiian Students	Vehicle Theft Index & Rates
	Percentage of American Indian Students	ZIP Code

Table 3*Personal Information Variables and Their Possible Values (If Categorical)*

Personal Information		
Extracurriculars	Home's Area	Distance to Schools
Participated in Cultural Activities	Area's Population	Distance Between Home & High School
Participated in Community Service	Size of the Area (ZIP Code)	Distance Between Home & College
	Assault Index & Rates	
	Burglary Index & Rates	
	Larceny Index & Rates	
	Murder Index & Rates	
	Personal Crime Index & Rates	
	Property Crime Index & Rates	
	Rape Index & Rates	
	Robbery Index & Rates	
	Total Crime Index & Rates	
	Vehicle Theft Index & Rates	
	ZIP Code	

Table 4*Social Demographic Variables and Their Possible Values (If Categorical)*

Social Demographic Information			
Gender	Ethnicity	Eligibility	Miscellaneous
Female	African American/Black	At Risk for Academic Failure	Had a Job While in High School?
Male	Asian	First Generation	Age At High School Graduation
	Hawaiian	First Generation & At High Risk for Academic Failure	
	Hispanic	Low Income	
	Multinational	Low Income & At High Risk for Academic Failure	
	White	Low Income & First Generation	
		Low Income, First Generation, & At High Risk for Academic Failure	

Before we began modeling, we set aside approximately fifteen percent of the dataset for model deployment. This dataset remained unused until we completed all modeling attempts. Reserving these twenty-two rows of unseen data allowed us to build our best anomaly detection model and simulate its efficacy when applied to a new dataset. We will demonstrate this process in the next chapter. We also divided the remaining “modeling” data into three groups to use with the supervised learning models (discussed later in this chapter). Specifically, we allotted seventy percent of the data (eighty-six students) for training, twenty percent (twenty-six students) for validation, and the final ten percent (twelve students) for testing. Supervised models also require their datasets to be separated into predictor and target variables. A target variable is the desired output or end result of a model; a predictor variable is the input for a model that is used to compute the target. In this project, the target variable is `PS_Graduated`, which indicates

whether the student graduated from or dropped out of college. We separated this column from the rest of the dataset, creating a target array (a single column of numbers). The remaining variables formed the predictor data matrix (several columns of numbers stacked side by side). Finally, we used Scikit-Learn's `StandardScaler` transformer to create standardized versions of all the data matrices.

Research Design

When designing this project, we understood it was impossible for the Upward Bound dataset to perfectly simulate every student's entire high school experience; not every student will attend an Upward Bound program. However, we strategically compiled and created our dataset's variables in an attempt to emulate that information. In general, we used enrollment in Upward Bound to represent administration providing support and assistance to students who are at risk for dropping out. We used the `Grade_EnteredUB` variable to denote which high school grade level a student first received support services in. Similarly, the `Participation` variable symbolized when a student received such services (i.e., if they were provided year-round, only during the academic year, only during the summer, etc.) Within the context of the Annual Performance Reports, the `CulturalAct` and `CommServ` variables track a student's involvement in Upward Bound's cultural and community service activities. Instead, we included these fields to represent a student's involvement in extracurricular activities during the school year. Finally, we incorporated the thirty-nine college dummy variables into the final dataset to illustrate a high school student expressing interest in attending that particular college.

To distinguish our research from other studies done within the field of education, we decided to employ both supervised and unsupervised anomaly detection techniques and compare their performances. With supervised learning tools, "each observation...[has] an associated

response measurement [or target]” (James, et al., 2013, p. 26). These models are “supervised” because the user structures how they work and identifies what the outcomes should be. Conversely, unsupervised learning tools treat each row as “a vector of measurements...[with] no associated response” (James, et al., p. 26). Instead of instructing these models on what they should do, users provide no structure, allowing the models to “seek to understand the relationships between the variables or between the observations” (James, et al., p. 27) and share their findings.

In Machine Learning, we use anomaly detection to identify rare “observations which...[differ] significantly from the majority of the data,” and there are “numerous anomaly detection methods” (Goyal, 2019). Since these instances exist outside of the norm, they are often referred to as outliers. For our project, we consider a student’s decision to drop out of school as an anomaly. To model this, we selected the three popular anomaly detection approaches most appropriate for this research: distance-based, clustering-based, and classification-based techniques.

Distance-Based Techniques

Distance-based techniques rely “on some measure of distance, or [a] distance-derived metric between points and sets of points” (Espressius, 2021). The Euclidean, Weighted Euclidean, Minkowski, Manhattan, and Mahalanobis distances are among some of the most commonly used distance equations (Espressius). The user selects one of these metrics and computes the distance between the points (or instances) within the dataset. The user also identifies a threshold value, and any points with distances larger than that threshold are labeled as outliers. This approach “[scales] well for large datasets with medium-to-high dimensionality;” however, the “high dimensionality drastically reduces performance due to elevated

computational complexity” (Espressius). In this project, we implemented all distance-based techniques as unsupervised Machine Learning tools.

Clustering-Based Techniques

Clustering-based techniques identify clusters within a dataset. These clusters are formed “such that objects [within them] are similar to each other” (Aravindharavindh, 2021). If there are any “points that are not within a cluster...[they are] considered anomalies” (Mehrotra, et al., 2017, p. 41). We can create these groups by implementing partition, hierarchical, density-based, or grid-based methods (Aravindharavindh, 2021). As with the distance-based approaches, users identify a threshold value to distinguish the anomalies from the clustered instances. Clustering models are able to “detect outliers without labeling the data,” (Aravindharavindh) making them excellent unsupervised learning strategies. However, their efficacy “largely depends on the clustering method used” (Aravindharavindh). This can make it very difficult and time-consuming to find the best technique for each dataset. In this project, we implemented all clustering-based techniques as unsupervised Machine Learning tools.

Classification-Based Techniques

Classification-based techniques use training data to fit “a classifier using the available labeled training data” (Editorial, 2021). This training data contains observations from at least two groups, more commonly referred to as classes. The model analyzes all instances within the training set and attempts to identify similarities within each class. When applied to testing data, the model uses these similarities to “[classify each] test instance as normal or abnormal” (Editorial). Since these models learn from the labels in the training data and assign a label to each unseen observation, they can only be used in supervised settings. These techniques “use powerful algorithms that can distinguish between instances belonging to different classes,”

(Editorial) often resulting in highly accurate models. However, if the model works “too hard to find patterns in the training data...[it] may be picking up some patterns that are just caused by random chance” (James, et al., 2013, p. 32). In data science, we use the term “overfitting” to refer to this phenomenon. In these instances, the model performs very poorly on unseen data, as the “patterns...found in the training data simply [do not] exist in the test data” (James, et al., p. 32). In this project, we implemented all classification-based techniques as supervised Machine Learning tools.

Assumptions

We conducted our research under the following assumptions:

- 1) The crime rates reported by Esri remain constant over time, with a small error margin. As such, their most recent dataset (from 2019) is reflective of the crime rates existing within our dataset’s twenty-one ZIP Codes from 2008 to 2015.
- 2) The information reported by the NCES and the NJDOE remain constant over time, with a small error margin. As such, their most recent datasets and reports (from the 2021 and 2020 school years, respectively) accurately depict the educational and demographic situations within our dataset’s ten high schools from 2008 to 2015.
- 3) Any factors the models identify as associated with dropping out of college are associated for the entire duration of our dataset (i.e., from 2008 to 2021). In other words, the associated factors do not fluctuate over time. As such, we can combine students from all seven cohorts into one dataset and apply the same modeling techniques to all students, irrespective of which year the student graduated high school.

Variable Selection

Before we could proceed with modeling, we needed to remove unnecessary variables from our dataset. In order to build effective models, it is important to only include “enough relevant features and not too many irrelevant ones” (Géron, 2019, p. 27). This will help to minimize overfitting (a concept discussed earlier in this section). We discovered that the suggested list of features to remove varied across different approaches. If we chose to drop the nine columns with either the same value for all students or those only used for the exploratory data analysis (detailed in the next section), we were left with one-hundred seventy-one columns in our dataset (Column Set A). If we chose to drop the five columns we suspected were redundant with other columns in the dataset along with the nine columns removed for Column Set A, we would obtain a dataset with only one-hundred sixty-six columns (Column Set B). If we chose to drop all forty-five columns found to be unassociated with the target variable (`PS_Graduated`) during our exploratory data analysis, our dataset would only have one-hundred thirty-three columns (Column Set C). In an attempt to build the most effective anomaly detection tool, we decided to build all of our models three times, once for each column set. As such, throughout the remainder of the modeling section, all tables will display a model’s performance on each column set. The comprehensive tables for Column Sets A, B, and C are provided in Appendices D, E, and F, respectively.

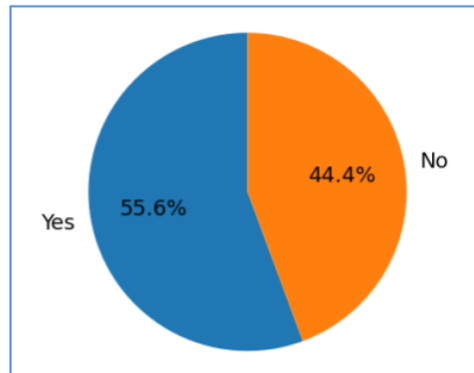
Exploratory Data Analysis

In data science, we often conduct an exploratory data analysis (EDA) “to analyze and investigate data sets and summarize their main characteristics” (IBM Cloud Education, 2020). This process allows us “to see what data can reveal beyond the formal modeling or hypothesis testing task and provides...a better understanding of [dataset] variables and the relationships

between them” (IBM Cloud Education). For this project, we applied various EDA techniques to

Figure 3

Dataset's Graduation Distribution



the modeling dataset to assess the distribution of each column, view the statistical summaries of all fifty continuous variables, and search for any correlations with the target variable and among the predictors. We used these correlations to determine which columns should be removed from the final dataset, as we previously alluded to. In this section, we highlight the most interesting

findings that appeared throughout this process.

Of the one-hundred twenty-four students in the dataset, sixty-nine graduated from college and the remaining fifty-five dropped out; the equivalent percentages are displayed in Figure 3. Surprisingly, these values also correspond to the dataset's gender distribution (Figure 4); however, through our analysis, we found thirty of the dropouts are females and the remaining twenty-five are males.

Hence, thirty-nine females and thirty males graduated. Figure 5 summarizes this information.

Figure 4

Dataset's Gender Distribution

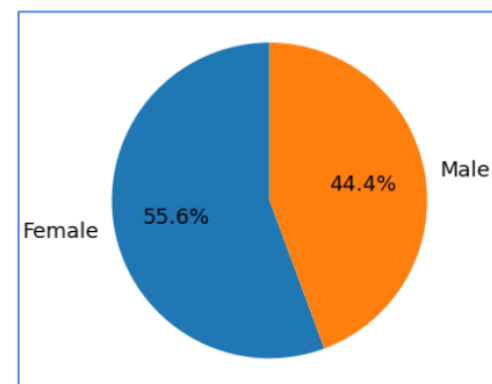
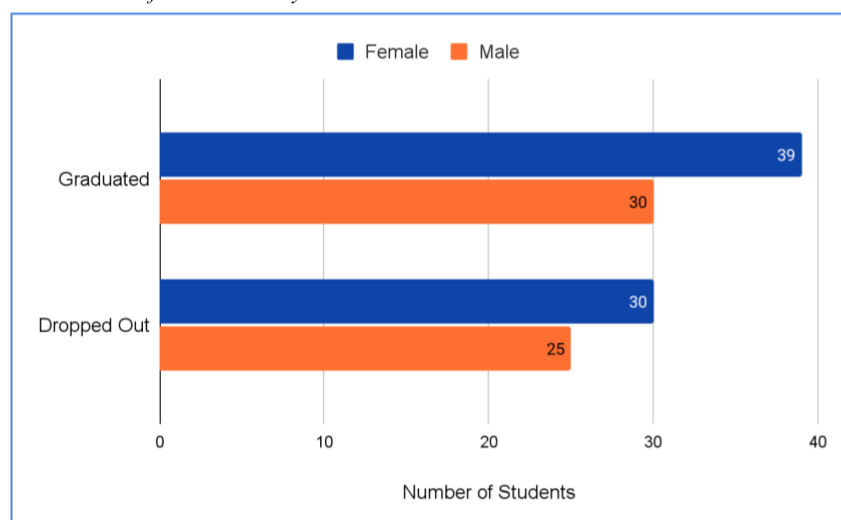


Figure 5

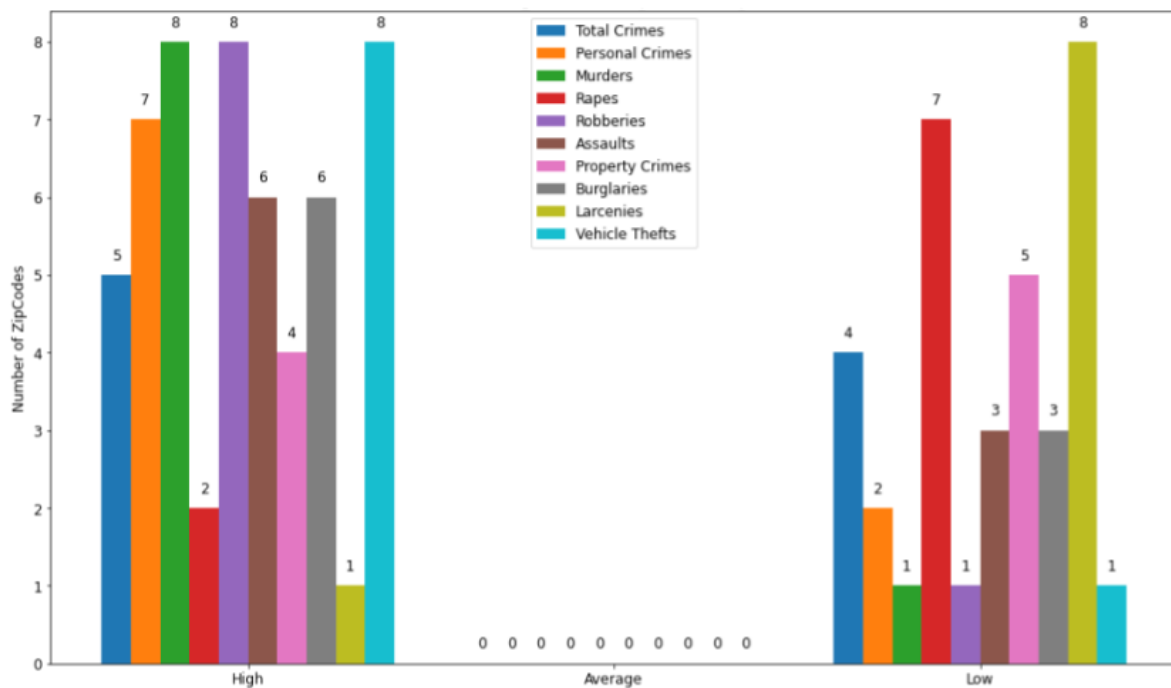
Distribution of Graduation by Gender



In Figure 6, we present the number of high school ZIP Codes that experience high, average, and low crime rates for each crime category. Most of the high schools are in areas with high murder, robbery, and vehicle theft rates. Conversely, almost all the schools' areas experience low larceny rates. By using a correlation matrix, we also found associations between some of the crime levels and specific groups of students. For example, our dataset indicates schools with higher percentages of students with disabilities are more likely to be in areas with high levels of personal crimes, murders, rapes, assaults, and vehicle thefts. Meanwhile, schools that have a larger number of students receiving free or reduced lunch are more likely to be in areas with low levels of burglaries.

Figure 6

Level of Crime in Each High School ZIP Code (9 ZIP Codes in Total)



Modeling

Throughout the modeling process, we used our twelve base models to build over one-hundred anomaly detection tools. In this section, we will highlight only a few of the results from each model. To view the outcomes of all our attempts, please refer to the tables in Appendices G–Q. Below, we provide a brief overview of how each model works and detail its performance within our project.

Modeling with Distance-Based Techniques

In our project, we implemented two unsupervised distance-based techniques: k -Nearest Neighbors and Pruning.

K -Nearest Neighbors (KNN). There are four possible ways to use k -Nearest Neighbors models for unsupervised learning, depending on which points are considered for the distances. The user indicates which points should be used for comparisons along with the number of neighboring points to find for each observation. When used in an unsupervised setting, “it is impossible to know the correct value for k for any particular algorithm, as this is highly dataset-dependent” (Espressius, 2021). As such, the list of outliers identified by the model is often different, depending on which method is used. In all situations, the algorithm iterates through each of the points and selects k of the remaining points to compute distances with. The four possible methods indicate which mathematical operation to perform on them. With the Distance to All Points method, the model calculates the sum of the k distances; meanwhile, the Distance to Nearest Neighbor approach only considers the minimum of these distances. With the Average and Median Distance methods, the model finds the average or median of the k distances, respectively. (Espressius).

For our research, we employed the Distance to Nearest Neighbor method; with this approach, a “point is considered anomalous if the distance to its nearest point is greater than some threshold” (Espressius, 2021). Throughout our attempts, we experimented with various k -values, thresholds, and distance metrics. We trained with both the scaled and unscaled datasets, oscillating between employing the complete set of columns and omitting the `PS_Graduated` variable. However, the maximum number of outliers found by a single model was fourteen (Table 5). Furthermore, collectively the nine models identified no more than twenty outliers; ten of these were dropouts. Thus, the k -Nearest Neighbors approach only identified approximately 20.00% of the college dropouts and approximately 18.18% with Column Sets B and C.

Table 5

Number of Outliers Identified by Various k -Nearest Neighbors Models

Number of Outliers With k -Nearest Neighbors			
Model	Column Set A	Column Set B	Column Set C
<u>Attempt #1</u> Full & Unscaled Dataset; $k = 3$	10	10	10
<u>Attempt #2</u> Full & Scaled Dataset; $k = 3$	14	12	11
<u>Attempt #4</u> Just Predictors & Scaled Dataset; $k = 3$	12	12	10
<u>Attempt #7</u> Just Predictors & Unscaled Dataset; $k = 2$	7	7	7
<u>Attempt #9</u> Just Predictors & Unscaled Dataset; $k = 4$	11	11	11
Across 9 Attempts	20	19	19

Pruning. Distance-based pruning methods are frequently referred to as “an extension” of k -Nearest Neighbors that prioritizes “a reduction in computational complexity” (Espressius, 2021). This is achieved by partitioning “the input space into discrete regions, with summary statistics such as the minimum-bounding rectangle” (Espressius). Instead of comparing a point with every remaining point, the algorithm compares it “to the bounding rectangle within which it lies [to] determine first if it is possible at all for a nearby region to contain [neighbors]. If not, the region nearby is eliminated” (Espressius). These pruning methods provided no improvement to the k -Nearest Neighbors models; instead, the pruned models were identical to those of the former. This duplication could be due to the relatively small size of our dataset.

The metrics used to assess a model’s performance are context-dependent. Because our project focuses on building a model to identify the high school students most likely to drop out of college, we need a model that can correctly identify college dropouts as outliers. Since these two techniques attained identical results (Table 6), we determined the k -Nearest Neighbors model was the best distance-based approach.

Table 6

Best Performance Attained by Each Distance-Based Model

Highest Percentage of College Dropouts Identified With Each Distance-Based Technique			
Model	Column Set A	Column Set B	Column Set C
k -Nearest Neighbors	20.00%	18.18%	18.18%
Pruning	20.00%	18.18%	18.18%

Modeling with Clustering-Based Techniques

Within our project, we implemented three clustering-based techniques: k -Means, Agglomerative/Hierarchical, and Gaussian Mixture Models. We trained all three in unsupervised settings. However, due to their clustering nature, we used supervised performance metrics to interpret their results.

K -Means. The k -Means algorithm attempts to identify a predetermined number of clusters within an unlabeled dataset. To do so, the model iterates through a two-step process until an indicated stopping point is reached. The model identifies k arbitrary points as the centroids; these are the centers of the clusters. Then, the model assigns each instance within the dataset “to the cluster whose centroid is closest” (Géron, 2019, p. 241). The model determines the new centroid values “by computing the mean of the instances for each cluster” (Géron, p. 241). Theoretically, this iterative process continues until the computed means are equal to the previous centroid values. However, users can choose to identify a maximum number of iterations or a maximum difference value to reduce a model’s runtime.

In the context of our research, two is the only appropriate value for k . We only care to distinguish between two groups of people: those who graduate from college and those who drop out. We would be unable to make any meaningful interpretations of models built with larger k -values. As such, all eight of our attempts utilized two clusters. Surprisingly, these eight models resulted in only two distinct lists of outliers. From Table 7, we see Attempts #1 and #3 only identified one student as a dropout, while the remaining attempts identified twenty. We built the better-performing models with the scaled dataset. These results are unsurprising, as k -Means models assume “all features are equally scaled” (Pramoditha, 2020). In Attempts # 1 and #3, the models correctly identified 0.00% of the dropouts, and the remaining attempts identified 20.00%.

Table 7

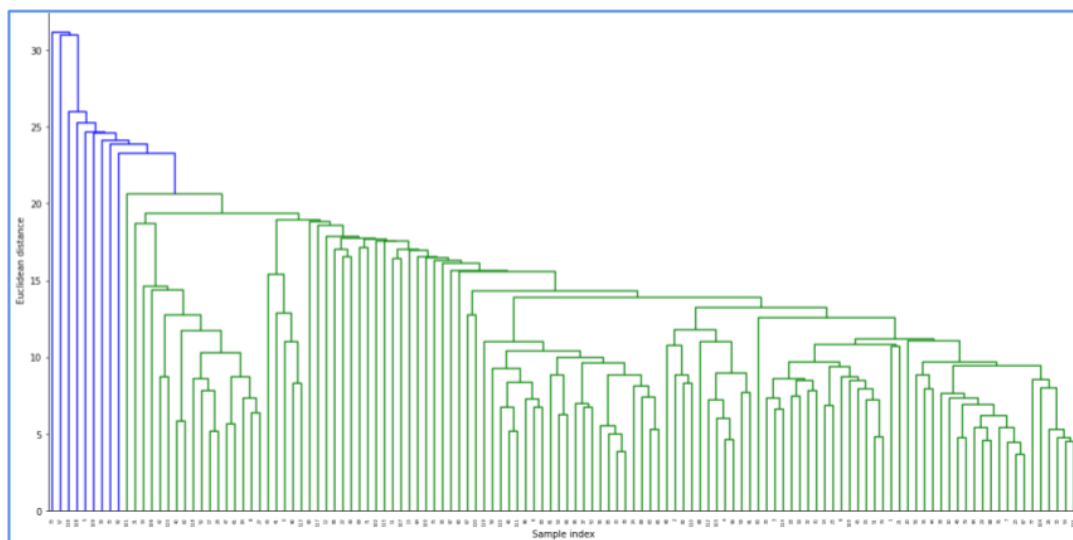
Number of Dropouts Predicted by Various k -Means Models

k-Means Models' Predicted Number of Dropouts			
Model	Column Set A	Column Set B	Column Set C
<u>Attempt #1</u> Full & Unscaled Dataset	1	1	1
<u>Attempt #2</u> Full & Scaled Dataset	20	20	20
<u>Attempt #3</u> Just Predictors & Unscaled Dataset	1	1	1
<u>Attempt #5</u> Full & Scaled Dataset; max_iter = 1000	20	20	20
<u>Attempt #7</u> Full & Scaled Dataset; n_init = 20	20	20	20

Agglomerative/Hierarchical. Unlike the previous anomaly detection models, hierarchical clustering “does not require that we commit to a particular choice of k ” (James, et al., 2013, p. 390). These models convey the clusters with a “tree-based representation of the observations, called a dendrogram;” (James, et al., p. 390) an example is shown in Figure 7.

Figure 7

Example of a Dendrogram



There are several variations of hierarchical clustering; in this project, we focused on the agglomerative approach. This technique is frequently referred to as “bottom-up clustering,” as it builds the dendrograms “starting from the leaves and combining clusters up to the trunk” (James, et al., p. 391). In other words, we first “take all data points as clusters and start merging[them] based on the distance between clusters;” we continue with this process “until we form one big cluster” (Ranjan, 2020). The heights at which the various clusters merge correspond to the similarity of the observations. The lower a merge occurs, “the more similar the groups of observations are to each other” (James, et al., 2013, p. 391).

Dissimilarities are determined by computing a distance metric on two clusters and applying the linkage attribute. Although there are several linkage types, “average, complete, and single linkage are most popular among statisticians” (James, et al., 2013, p. 394). These metrics utilize the average, maximal, and minimal intercluster distances, respectively, between all pairs (James, et al., p. 395). Once the dendrogram is completed, we can identify clusters by drawing a horizontal line at a selected vertical cutoff point. As such, “the height of the cut to the dendrogram serves the same role as the k in k -Means clustering: it controls the number of clusters obtained” (James, et al., p. 393).

While building our agglomerative models, we employed all three of the aforementioned linkage types. We also experimented with various distance metrics. To select the vertical cutoff points, we followed the same line of reasoning as we did when selecting k in k -Means; all cuts had to be drawn so the dendrogram only contained two clusters. Like the best k -Means models, the most effective agglomerative models trained on the scaled dataset, as indicated in Table 8. In most of the attempts, these hierarchical models correctly identified 0.00% of the dropouts; however, the best models correctly identified approximately 41.82% of those students.

Table 8

Number of Dropouts Predicted by Various Agglomerative/Hierarchical Models

Agglomerative Models' Predicted Number of Dropouts			
Model	Column Set A	Column Set B	Column Set C
<u>Attempt #1</u> Full & Unscaled Dataset; affinity = "euclidean"; linkage = "complete"	1	1	1
<u>Attempt #12</u> Full & Unscaled Dataset; affinity = "cosine"; linkage = "average"	33	33	33
<u>Attempt #13</u> Full & Scaled Dataset; affinity = "cosine"; linkage = "average"	45	45	45
<u>Attempt #14</u> Full & Unscaled Dataset; affinity = "cosine"; linkage = "single"	7	7	7
<u>Attempt #17</u> Full & Scaled Dataset; affinity = "cosine"; linkage = "complete"	48	45	44

Gaussian Mixture Models (GMM). A Gaussian Mixture model is a “probabilistic model that assumes all the data points are generated from a mixture of a finite number of gaussian distributions” (Gumbao, 2019). During training, these models “try to recover the original gaussian that generated this distribution” (Gumbao). As with the k -Means models, users must select a value for k before modeling with Gaussian Mixtures. In fact, the process for building these models is very similar to that of k -Means. A Gaussian Mixture model “initializes the cluster parameters randomly, then it repeats two steps until convergence, first assigning instances to clusters...and then updating the clusters” (Géron, 2019, p. 262). This algorithm is considered to be “a generalization of k -Means,” as it “finds the cluster centers...[along with]

their size, shape, and orientation...as well as their relative weights” (Géron, p. 262). Once all the observations are clustered, these models refer to the low-density regions to detect anomalies. Just like the distance-based techniques, Gaussian Mixture models require users to “define what density threshold [they] want to use” (Géron, p. 266). The models use these predetermined values to identify the appropriate percentage of outliers within the low-density regions.

Typically, these thresholds should reflect the “real world” trends. In our models, we set our density threshold to thirty-three percent, as research shows one-third of college students drop out and “never earn a degree” (Leonhardt & Chinoy, 2019). Continuing the line of reasoning from our other clustering-based models, we indicated all Gaussian Mixture models should have two clusters. As such, there was very little we could do to tune these models. In fact, we only conducted four attempts, all of which are displayed in Table 9, below. Unsurprisingly, the best Gaussian Mixture models trained on the scaled modeling data. All twelve models (four with each column set) correctly identified at least 30.00% of the dropouts, and the best attempts identified approximately 69.09%. However, the models trained with Column Set C identified a maximum of 32.73% of the college dropouts.

Table 9

Number of Dropouts Predicted by the Gaussian Mixture Models

Gaussian Mixtures’ Predicted Number of Dropouts			
Model	Column Set A	Column Set B	Column Set C
<u>Attempt #1</u> Full & Unscaled Dataset	31	31	31
<u>Attempt #2</u> Full & Scaled Dataset	68	68	32
<u>Attempt #3</u> Just Predictors & Unscaled Dataset	31	31	31
<u>Attempt #4</u> Just Predictors & Scaled Dataset	66	68	32

As a reminder, we want to prioritize models that correctly identify the largest percentage of college dropouts. Using the results presented in Table 10, we decided the Gaussian Mixture approach produced the best clustering-based anomaly detection tools.

Table 10

Best Performance Attained by Each Clustering-Based Model

Highest Percentage of College Dropouts Identified With Each Clustering-Based Technique			
Model	Column Set A	Column Set B	Column Set C
<i>k</i> -Means	20.00%	20.00%	20.00%
Agglomerative/Hierarchical	41.82%	41.82%	41.82%
Gaussian Mixture Models	69.09%	69.09%	32.73%

Modeling with Classification-Based Techniques

In our project, we implemented seven supervised classification-based techniques: Stochastic Gradient Descent, Logistic Regression, Support Vector Machines, Decision Trees, Random Forests, Boosting, and Voting Classifiers. Due to similarities in their algorithms, these four techniques can be segmented into four distinct categories.

Stochastic Gradient Descent (SGD) and Logistic Regression. In Machine Learning, we often use a Gradient Descent algorithm to find “optimal solutions,” as this technique modifies various parameters as it iterates through a dataset “in order to minimize a cost function” (Géron, 2019, p. 118). In other words, this algorithm uses information directly from the training data to tune its parameters in an effort to achieve the smallest possible error. To achieve this, the model calculates “the gradient of the cost function with regard to each model parameter θ_j ” (Géron, p.

121) using Equation (1). In mathematics, this notion is more commonly referred to as the partial derivative.

$$\frac{\delta}{\delta \theta_j} \text{MSE}(\theta) = \frac{2}{m} \sum_{i=1}^m (\theta^T \mathbf{x}^{(i)} - y^{(i)}) x_j^{(i)} \quad (1)$$

There are three variations of this Gradient Descent method. With Batch Gradient Descent, the model computes all the partial derivatives at once (Géron, 2019, p. 121). With Stochastic Gradient Descent, the model “picks a random instance in the training set at every step and computes the gradients based only on that single instance” (Géron, p. 124). Lastly, in Mini-batch Gradient Descent, gradients are computed on “small random sets of instances” (Géron, p. 127).

For our research, we implemented Stochastic Gradient Descent. When used as a classification tool, the model trains “instances independently, one at a time” (Géron, 2019, p. 88). In addition to building numerous Stochastic Gradient Descent models, we used Scikit-Learn’s `SGDClassifier` to build Logistic Regression models. Logistic Regression is “used to estimate the probability that an instance belongs to a particular class...If the estimated probability is greater than 50%, then the model predicts that the instance belongs to that class...and otherwise it predicts that it does not” (Géron, p. 142). Throughout our fifteen attempts, we implemented various penalty and alpha metrics and trained with both the scaled and unscaled datasets. Since data scientists suggest users “ensure that all features have a similar scale” before building Gradient Descent models (Géron, 2019, p. 121), it is unsurprising that almost all of the better-performing models trained with the scaled dataset (Table 11).

Unlike the models from the distance-based and clustering-based techniques, the performance rates of these models were very different across the three column sets. The models built with Column Set A identified a minimum of 50.00% of the college dropouts, with the best

performing model correctly identifying approximately 85.71% of the students. The predictive abilities of the models built with Column Set B ranged from approximately 35.71% to approximately 78.57%. The best performing model built with Column Set C also correctly identified approximately 78.57% of the college dropouts, though the corresponding worst performing model was slightly more accurate at approximately 42.86%. Throughout the remainder of this chapter, we will see this phenomenon is not unique to these two models; rather, it is something that remains consistent across all classification-based techniques. In Chapter Four, we will analyze the models to determine the cause of the discrepancies.

Table 11

Number of Dropouts Predicted by Various Stochastic Gradient Descent and Logistic Regression Models

Stochastic Gradient Descent and Logistic Regression Models' Predicted Number of Dropouts			
Model	Column Set A	Column Set B	Column Set C
<u>Attempt #1</u> Unscaled Dataset	20	20	20
<u>Attempt #2</u> Scaled Dataset	13	11	12
<u>Attempt #6</u> Scaled Dataset; penalty = "l2"; alpha = 0.01	11	10	19
<u>Attempt #8</u> Scaled Dataset; loss = "log"	14	12	12
<u>Attempt #10</u> Scaled Dataset; loss = "perceptron"	13	11	12

Support Vector Machines (SVM). A Support Vector Machine is a “versatile Machine Learning model capable of performing linear or nonlinear classification, regression, and...outlier detection” (Géron, 2019, p. 153). When used for classification, Support Vector Machines utilize

a special function, called a kernel, which “quantifies the similarity of two observations” (James, et al., 2013, p. 352). Implementing these kernels help the models group like instances together. To distinguish between the various classes, these models use boundary lines (i.e., hyperplanes) to divide the sample space. These classifiers use this information to “not only [separate] the two classes but also [stay] as far away from the closest training instances as possible” (Géron, 2019, p. 153). There are several linear and nonlinear kernel functions, but linear, polynomial, and radial are among the most popular (James, et al., 2013, p. 352). In all three situations, the models seek “to find a good balance between keeping the [margin] as large as possible” while also minimizing the number of misclassified observations (Géron, 2019, p. 155).

Table 12

Number of Dropouts Predicted by Various Support Vector Machine Models

Support Vector Machine Models’ Predicted Number of Dropouts			
Model	Column Set A	Column Set B	Column Set C
<u>Attempt #1</u> Unscaled Dataset; StandardScaler; LinearSVC; loss = “hinge”	15	11	13
<u>Attempt #3</u> Unscaled Dataset; StandardScaler; LinearSVC; loss = “squared_hinge”	15	11	15
<u>Attempt #8</u> Scaled Dataset; SVC; gamma = “scale”	7	8	5
<u>Attempt #11</u> Unscaled Dataset; PolynomialFeatures; StandardScaler; LinearSVC; loss = “squared_hinge”	14	4	12
<u>Attempt #15</u> Unscaled Dataset; StandardScaler; SVC; kernel = “rbf”; gamma = 5	0	8	0

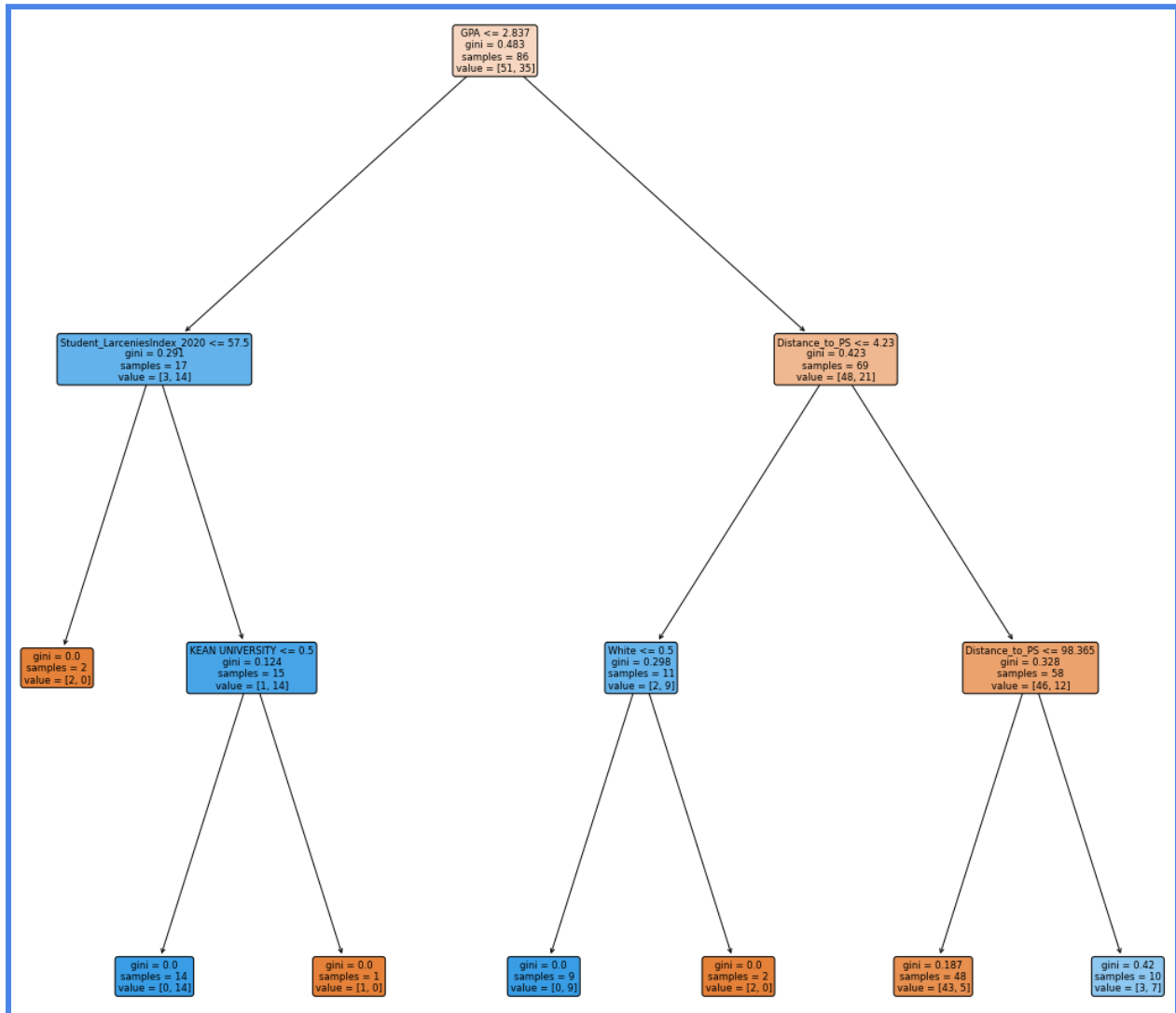
We built sixteen models with each column set, experimenting with various combinations of kernel functions and scaled or unscaled datasets. In Column Sets A and C, models using a linear kernel provided the best performing models, correctly identifying approximately 85.71% and 57.14% of the college dropouts, respectively. Meanwhile, the best performing model of Column Set B, which used the Gaussian Radial Basis Function kernel, only identified approximately 35.71% of the dropouts. The results of five of our attempts, along with the parameters used to tune them, are displayed in Table 12.

Decision Trees, Random Forests, and Boosting. Decision Trees (specifically, Classification Trees) “[segment] the predictor space into a number of simple regions” (James, et al., 2013, p. 303) and predicts “that each observation belongs to the most commonly occurring class of training observations” within that region (p. 311). These models are referred to as trees because we can depict their “[sets] of splitting rules used to segment the predictor space” (James, et al., p. 303) with branching structures like the one shown in Figure 8. Each splitting node contains an inequality expression. To classify an instance, the model begins at the starting node and evaluates the inequality for that observation. If the expression is true for that observation, the model continues down the left branch of the tree, and if not, the model proceeds down the right branch of the tree. This process continues until a terminal node is reached. This bottom node indicates which class the observation is predicted to be in. Due to their intuitive and simplistic nature, these models are very easy to understand. Unfortunately, they are often “not competitive with the best supervised learning approaches...in terms of prediction accuracy” (James, et al., p. 303). Consequently, statisticians developed extensions of Decision Trees, such as Random Forests and Boosting, which “[involve] producing multiple trees which are then combined to yield a single consensus prediction” (James, et al., p. 303). Oftentimes, “combining a large

number of trees can often result in dramatic improvements in prediction accuracy, at the expense of some loss in interpretation” (James, et al., p. 303).

Figure 8

Example of a Decision Tree



Random Forest models “build a number of Decision Trees on bootstrapped training samples” (James, et al., 2013, p. 319). This means the model resamples observations from the training set to create “new” datasets. This approach incorporates a technique that “decorrelates the trees” (James, et al., p. 319). Specifically:

When building these Decision Trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors. A fresh sample of m predictors is taken at each split, and typically we choose $m \approx \sqrt{p}$. (James, et al., p. 319)

By restricting the number of features available as split candidates in this manner, “the algorithm is not even allowed to consider a majority of the available predictors” (James, et al., p. 319). Typically, this increase in tree diversity reduces the variance and produces better performing models.

Boosting algorithms build their models by sequentially training on the data. Essentially, “each tree is grown using information from previously grown trees” (James, et al., 2013, p. 321). Unlike Random Forest models, “Boosting does not involve bootstrap sampling; instead each tree is fit on a modified version of the original [dataset]” (James, et al., p. 321). Although these models start with a weak learner, they sequentially improve from previous attempts until they build a strong learner. There are numerous Boosting methods, including AdaBoost and Gradient Boosting. The former prioritizes learning from “the training instances that the predecessor underfitted” (Géron, 2019, p. 200), while the latter “tries to fit the new predictor to the residual errors made by the previous predictor” (p. 203). Within our research, we implemented AdaBoost.

Throughout our research, we built numerous models using each of these three tree-based techniques. In fact, we fit seven, eight, and sixteen models, respectively, with each column set. To avoid overfitting, we employed several pruning techniques. Such approaches included limiting the tree’s maximum depth and restricting the number of instances required to split or terminate a node. As shown in Table 13, the three approaches demonstrated very similar predictive abilities within each column set.

Table 13*Best and Worst Performance Attained by Each Tree-Based Technique*

Highest & Lowest Percentage of College Dropouts Identified With Each Tree-Based Technique			
Decision Tree			
Metric	Column Set A	Column Set B	Column Set C
Highest	100.00%	42.86%	50.00%
Lowest	64.29%	28.57%	35.71%
Random Forest			
Metric	Column Set A	Column Set B	Column Set C
Highest	78.57%	42.86%	64.29%
Lowest	64.29%	28.57%	35.71%
Boosting			
Metric	Column Set A	Column Set B	Column Set C
Highest	100.00%	57.14%	57.14%
Lowest	57.14%	28.57%	35.71%

Voting Classifiers. In Machine Learning, it is good practice to “aggregate the predictions of a group of predictors” using an ensemble learning method, as it “will often get better predictions than with the best individual predictor” (Géron, 2019, p. 189). The Random Forest and Boosting algorithms are among “the most popular Ensemble methods” (Géron, p. 189); however, these approaches only aggregate the predictions of Decision Trees. Meanwhile, the Voting Classifier technique is able to compile the predictions of a diverse set of classifiers. In fact, a more diverse set of models typically attains higher accuracy. When classifiers are trained “using very different algorithms...[it] increases the chance that they will make very different types of errors, improving the ensemble’s accuracy” (Géron, p. 191). When Voting Classifiers “aggregate the predictions of each classifier and predict the class that gets the most votes,” the

Ensemble is fit as a hard voting classifier (Géron, p. 190). In doing so, the classifier leverages the fact that all the models have different approaches to classifying observations.

We decided to train the Voting Classifier using the best-performing model from each of the three aforementioned classification-based techniques. These three models differed among the three column sets. With Column Set A, the best models were a Logistic Regression implementation of Stochastic Gradient Descent, a Support Vector Machine with a linear kernel, and a Boosting model. With Column Set B, the best models were a Stochastic Gradient Descent model, a Support Vector Machine with a Gaussian Radial Basis Function kernel, and a Boosting model. With Column Set C, the best models were a Stochastic Gradient Descent model, a Support Vector Machine with a linear kernel, and a Random Forest model. Table 14 provides a summary of the performances of these three Voting Classifier models. These models correctly identified 100.00%, 78.57%, and 64.29% of the college dropouts, respectively.

Table 14

Number of Dropouts Predicted by Each Voting Classifier Model

Voting Classifiers' Predicted Number of Dropouts		
Column Set A	Column Set B	Column Set C
15	17	17

As a reminder, we want to prioritize models that correctly identify the largest percentage of college dropouts. We believe it is better to accidentally provide additional support to students who are unlikely to drop out of college than it is to not provide support to a student who most likely would. Therefore, we want to identify as many of the at-risk students as possible. Within our models, the positive instances corresponded to these students. Since the recall metric indicates the proportion of positive instances correctly detected by a classifier, the best models

are those with the highest recall values. Using the results presented in Table 15, we decided the Boosting, Voting Classifiers, and Stochastic Gradient Descent produced the best classification-based anomaly detection tools, depending on the column set.

Table 15

Best Performance Attained by Each Classification-Based Model

Highest Recall Value Attained With Each Classification-Based Technique			
Model	Column Set A	Column Set B	Column Set C
Logistic Regression	0.5000	0.2857	0.3571
Stochastic Gradient Descent	0.8571	0.7857	0.7857
Support Vector Machine	0.8571	0.3571	0.5714
Decision Tree	1.0000	0.4286	0.5000
Random Forest	0.7857	0.4286	0.6429
Boosting	1.0000	0.5714	0.5714
Voting Classifier	1.0000	0.7857	0.6429

Throughout this chapter, we have detailed how we collected, cleaned, and utilized data from five various sources to address our research question. We also provided an in-depth understanding of how we designed, conducted, and assessed our anomaly detection models. In the next chapter, we will discuss how we can narrow down our list of best-performing models and identify which technique produces the best early detection tool.

Chapter Four: Analysis and Discussion

In Chapter Three, we reviewed the various attempts of our twelve modeling methods to determine the best-performing model for each anomaly detection technique. In this chapter, we assess these three techniques to determine which is most appropriate for addressing college dropout rates. We will conduct an analysis of each model's numerical performance measures and textual characteristic findings. Assessing the characteristic findings will reveal if our models support the qualitative and quantitative studies presented in Chapter Two and will help us identify which factors are strongly associated with dropping out of college. We will then use the best technique to select the final models, fit them to the union of the training and validation datasets, and apply them to the testing dataset. Finally, we will decide which of the three column sets is most suitable and apply its corresponding model to the model deployment dataset.

Analysis

To determine which model is most effective for identifying students likely to drop out of college, we must compare the best-performing models of each technique. In Chapter Three, we found k -Nearest Neighbors were the best distance-based models; Gaussian Mixtures were the best clustering-based models; and Boosting, Voting Classifiers, and Stochastic Gradient Descent were the best classification-based models, depending on the column set. After exploring the various performance measures commonly used to assess a model's quality, we deemed a model's ability to correctly identify college dropouts as the most appropriate metric for our problem. This corresponds to the percentage of dropouts identified by unsupervised models and the recall value obtained by supervised approaches. In both scenarios, higher values indicate better models. Since Gaussian Mixture Models are compatible with both density-based and predictive methods, we

can assess its performance as both an unsupervised and a supervised anomaly detection tool. For simplicity, we will compare models of the same learning type, beginning with unsupervised.

Tables 16 and 17 summarize the performances of the best k -Nearest Neighbors and Gaussian Mixture models, respectively. Comparing each model's percentage of identified dropouts to its percentage of identified graduates reveals both techniques are most accurate when detecting dropouts. However, the k -Nearest Neighbors' highest percentage of identified dropouts

Table 16

Best Distance-Based Technique's Performance as an Anomaly Detection Tool

Best Distance-Based Technique's Performance			
Metric	Column Set A (k-Nearest Neighbors)	Column Set B (k-Nearest Neighbors)	Column Set C (k-Nearest Neighbors)
Number of Outliers	20	19	19
Number of Dropouts	11	10	10
Number of Graduates	9	9	9
% of Dropouts Identified	20.00%	18.18%	18.18%
% of Graduates Identified	13.04%	13.04%	13.04%

Table 17

Best Clustering-Based Technique's Performance as an Unsupervised Anomaly Detection Tool

Best Clustering-Based Technique's Performance			
Unsupervised			
Metric	Column Set A (Gaussian Mixture)	Column Set B (Gaussian Mixture)	Column Set C (Gaussian Mixture)
Number of Outliers	65	64	43
Number of Dropouts	38	38	21
Number of Graduates	27	26	22
% of Dropouts Identified	69.09%	69.09%	38.18%
% of Graduates Identified	39.13%	37.68%	31.88%

is twenty percent, which is only seven percent more than its percentage of identified graduates. This indicates the method has relatively equal predictive abilities across both groups. Furthermore, these low results suggest these models are very ineffective for our problem. When given a set of one hundred students who would drop out of college, the best k -Nearest Neighbors model would correctly identify only twenty of those students. In other words, simply assuming *all* students will drop out would be more accurate than the best distance-based approach.

Meanwhile, the Gaussian Mixture models' identified dropout percentages are significantly higher than their corresponding graduate percentages, further emphasizing that these models are most useful for determining which students will drop out of college (i.e., the primary goal of our research). The models trained with Column Sets A and B are more effective than the model trained on Column Set C; the former models are almost two times more accurate. When given a set of one hundred students who would drop out of college, these models would correctly identify sixty-nine of the students. Even so, the third Gaussian Mixture model's dropout performance is still approximately twenty percent more accurate than the k -Nearest Neighbors models. In the unsupervised setting, the Gaussian Mixture models outperform those produced by k -Nearest Neighbors, making clustering-based models the better choice for identifying college dropouts. We can attribute this to the customizable nature of the Gaussian Mixture models. As we mentioned in Chapter Three, the only two parameters we can tune in k -Nearest Neighbors models are the number of neighboring points to use and how to compare their distances. Neither of these have an obvious correlation with college dropout rates, making it difficult to tailor the models to our specific problem. Meanwhile, we can set the number of clusters in the Gaussian Mixture models to the number of groups we know exist within the dataset. In this context, we know there are students who dropped out of college as well as students who graduated; hence, it

makes sense to indicate there should be two clusters. Similarly, we can tailor the density threshold to mirror the trends we know exist in the data. We know approximately one-third of college students never earn a degree (Leonhardt & Chinoy, 2019); using a thirty-three percent threshold reflects this information.

Although these two techniques greatly differ in their numerical results, many of their important features overlap. Table 18 displays the student-centered characteristics each k -Nearest Neighbors model deemed to be most common among dropouts. We consider the factors that appear in all three column sets (highlighted in orange) to be most associated with a

Table 18

Best Distance-Based Technique's Characteristics of Students Most Likely to Drop Out

Best Distance-Based Technique's Findings: Student Characteristics		
Students With the Following Characteristics Are Likely to <u>DROP OUT</u>		
Column Set A (k-Nearest Neighbors)	Column Set B (k-Nearest Neighbors)	Column Set C (k-Nearest Neighbors)
African American/Black	African American/Black	African American/Black
Attends College Farther From Home	Attends College Farther From Home	Attends College Farther From Home
Attends High School Farther From Home	Attends High School Farther From Home	Attends High School Farther From Home
Enters Upward Bound in 9th Grade	Enters Upward Bound in 9th Grade	–
Female	Female	–
Home Area Has Higher Murder Rates	Home Area Has Higher Murder Rates	–
Home Area Has Higher Vehicle Theft Rates	Home Area Has Higher Vehicle Theft Rates	Home Area Has Higher Vehicle Theft Rates
Lives in a Larger Area	Lives in a Larger Area	Lives in a Larger Area
Lives in a More Populous Area	Lives in a More Populous Area	Lives in a More Populous Area
More Absences	More Absences	More Absences
Participates in Community Service	Participates in Community Service	–
Participates in Cultural Activities	–	–

student dropping out of college. From this table, we see only one of the seven important predictors is academic related: the number of absences. Instead, if we refer back to Tables 3 and 4 on pages 19 and 20, we see the distance-based approach indicates a student's personal and social demographic situations have the largest impact on college graduation status.

Five of these characteristics also appear in the Gaussian Mixture table (Table 19). Both the distance-based and clustering-based techniques indicate traveling farther distances between

Table 19

Best Clustering-Based Technique's Unsupervised Characteristics of Students Most Likely to Drop Out

Best Clustering-Based Technique's Unsupervised Findings: Student Characteristics		
Students With the Following Characteristics Are Likely to <u>DROP OUT</u>		
Column Set A (Gaussian Mixture)	Column Set B (Gaussian Mixture)	Column Set C (Gaussian Mixture)
African American/Black	African American/Black	–
Attends College Farther From Home	Attends College Farther From Home	Attends College Farther From Home
Attends High School Farther From Home	Attends High School Farther From Home	Attends High School Farther From Home
–	–	Home Area Has Higher Assault Rates
Home Area Has Higher Murder Rates	Home Area Has Higher Murder Rates	Home Area Has Higher Murder Rates
–	–	Home Area Has Higher Personal Crime Rates
–	–	Home Area Has Higher Rape Rates
–	–	Home Area Has Higher Robbery Rates
Home Area Has Higher Vehicle Theft Rates	Home Area Has Higher Vehicle Theft Rates	Home Area Has Higher Vehicle Theft Rates
Lives in a Larger Area	Lives in a Larger Area	Lives in a Larger Area
–	Lives in a Low Larceny Area	–
–	–	Lives in a More Populous Area
Low Income and First Generation	–	–
More Absences	More Absences	More Absences

home and school, earning repeated absences, living in larger neighborhoods, and living in areas with high vehicle theft rates are common characteristics of college dropouts. Because these five characteristics appear in all six of these unsupervised learning models, we can infer that students experiencing a combination of these situations are at increased risk for dropping out. Conversely, it is inconclusive if being African American, living in a highly populated area, and living in an area with other high crime rates affect a student's likelihood of dropping out.

Similarly, we can use Tables 20 and 21 to assess which characteristics are most frequently associated with the high schools of postsecondary dropouts. Collectively, the two techniques identify ten institutional characteristics. Their six overlapping traits suggest the area's crime rates and the student body's demographics are the most pertinent features. It is important to emphasize this demographic information is not student-specific. For example, although the *k*-Nearest Neighbors models identify schools with a large population of homeless students as more likely to educate individuals who will drop out, this does not mean the homeless students will be the dropouts; such traits merely describe the make-up of the student population.

Of all ten crime categories, only vehicle theft rates appear with both student and high school characteristics. Students who are exposed to elevated vehicle theft rates, either around home or school, are more likely to drop out of college. However, we do not have enough information to conclude whether this risk can be compounded. With these unsupervised methods, we do not have a numeric depiction of the level of association for each characteristic. As such, we cannot determine whether students who both live in and attend school in areas with high vehicle theft rates are more likely to drop out of college than those who exclusively live in or attend school in such areas.

Table 20

Best Distance-Based Technique's Characteristics of High Schools Most Likely to Create Dropouts

Best Distance-Based Technique's Findings: High School Characteristics		
High Schools With the Following Characteristics Are Likely to Create <u>DROPOUTS</u>		
Column Set A (<i>k</i>-Nearest Neighbors)	Column Set B (<i>k</i>-Nearest Neighbors)	Column Set C (<i>k</i>-Nearest Neighbors)
–	–	Area Has Higher Assault Rates
Area Has Higher Murder Rates	Area Has Higher Murder Rates	Area Has Higher Murder Rates
Area Has Higher Personal Crime Rates	Area Has Higher Personal Crime Rates	Area Has Higher Personal Crime Rates
Area Has Higher Rape Rates	Area Has Higher Rape Rates	Area Has Higher Rape Rates
Area Has Higher Robbery Rates	Area Has Higher Robbery Rates	Area Has Higher Robbery Rates
Area Has Higher Vehicle Theft Rates	Area Has Higher Vehicle Theft Rates	Area Has Higher Vehicle Theft Rates
–	In a More Populous Area	–
More African American/Black Students	More African American/Black Students	More African American/Black Students
–	–	More American Indian/Alaskan Native Students
More Homeless Students	More Homeless Students	More Homeless Students
–	–	More Native Hawaiian/Pacific Islander Students
More Students in Foster Care	More Students in Foster Care	More Students in Foster Care
–	–	More Students w/ Disabilities
–	More Students w/ Free Lunch	–
More Students w/ Reduced Lunch	More Students w/ Reduced Lunch	–
More White Students	More White Students	–
Title I School-Wide	Title I School-Wide	–

Table 21

Best Clustering-Based Technique's Unsupervised Characteristics of High Schools Most Likely to Create Dropouts

Best Clustering-Based Technique's Unsupervised Findings: High School Characteristics		
High Schools With the Following Characteristics Are Likely to Create <u>DROPOUTS</u>		
Column Set A (Gaussian Mixture)	Column Set B (Gaussian Mixture)	Column Set C (Gaussian Mixture)
Area Has Higher Assault Rates	Area Has Higher Assault Rates	Area Has Higher Assault Rates
–	–	Area Has Higher Crime Rates
Area Has Higher Murder Rates	Area Has Higher Murder Rates	Area Has Higher Murder Rates
Area Has Higher Personal Crime Rates	Area Has Higher Personal Crime Rates	Area Has Higher Personal Crime Rates
–	–	Area Has Higher Rape Rates
Area Has Higher Robbery Rates	Area Has Higher Robbery Rates	Area Has Higher Robbery Rates
Area Has Higher Vehicle Theft Rates	Area Has Higher Vehicle Theft Rates	Area Has Higher Vehicle Theft Rates
–	In a High Murder Area	–
–	In a High Robbery Area	–
–	In a High Vehicle Theft Area	–
–	–	In a Larger Area
–	In a More Populous Area	In a More Populous Area
More African American/Black Students	More African American/Black Students	More African American/Black Students
–	–	More American Indian/Alaskan Native Students
–	–	More Homeless Students
–	–	More Native Hawaiian/Pacific Islander Students
More Students in Foster Care	More Students in Foster Care	More Students in Foster Care
	More Students w/ Free Lunch	–
More Students w/ Reduced Lunch	More Students w/ Reduced Lunch	More Students w/ Reduced Lunch
More White Students	More White Students	–
	Title I	–
Title I School-Wide	Title I School-Wide	–

Since the distance-based and clustering-based models inadequately identify graduates, it is not surprising the two techniques have no overlapping traits pertaining to college graduates. In fact, the clustering-based technique provides no student characteristics associated with graduates (Table 22); only Column Set C identifies a “meaningful” association. Furthermore, this single association is one that the distance-based technique indicates as a characteristic of dropouts. In Table 23, we see the distance-based models’ two features are very similar crimes; thus, it is possible the two are correlated. Finally, both techniques suggest the former high schools of college graduates frequently have larger Asian student populations (Tables 24 and 25).

Table 22

Best Clustering-Based Technique’s Unsupervised Characteristics of Students Most Likely to Graduate

Best Clustering-Based Technique’s Unsupervised Findings: Student Characteristics		
Students With the Following Characteristics Are Likely to <u>GRADUATE</u>		
Column Set A (Gaussian Mixture)	Column Set B (Gaussian Mixture)	Column Set C (Gaussian Mixture)
–	–	Attends High School Farther From Home

Table 23

Best Distance-Based Technique’s Characteristics of Students Most Likely to Graduate

Best Distance-Based Technique’s Findings: Student Characteristics		
Students With the Following Characteristics Are Likely to <u>GRADUATE</u>		
Column Set A (<i>k</i>-Nearest Neighbors)	Column Set B (<i>k</i>-Nearest Neighbors)	Column Set C (<i>k</i>-Nearest Neighbors)
Home Area Has Higher Burglary Rates	Home Area Has Higher Burglary Rates	Home Area Has Higher Burglary Rates
Home Area Has Higher Robbery Rates	Home Area Has Higher Robbery Rates	Home Area Has Higher Robbery Rates

Table 24

Best Distance-Based Technique's Characteristics of High Schools Most Likely to Create Graduates

Best Distance-Based Technique's Findings: High School Characteristics		
High Schools With the Following Characteristics Are Likely to Create <u>GRADUATES</u>		
Column Set A (<i>k</i>-Nearest Neighbors)	Column Set B (<i>k</i>-Nearest Neighbors)	Column Set C (<i>k</i>-Nearest Neighbors)
–	Area Has Higher Burglary Rates	–
In a Larger Area	In a Larger Area	In a Larger Area
More Asian Students	More Asian Students	More Asian Students
More Hispanic Students	More Hispanic Students	More Hispanic Students

Table 25

Best Clustering-Based Technique's Unsupervised Characteristics of High Schools Most Likely to Create Graduates

Best Clustering-Based Technique's Unsupervised Findings: High School Characteristics		
High Schools With the Following Characteristics Are Likely to Create <u>GRADUATES</u>		
Column Set A (Gaussian Mixture)	Column Set B (Gaussian Mixture)	Column Set C (Gaussian Mixture)
–	–	Attends H.S. Farther From Home
–	–	Is in the 07502 ZIP Code
–	–	Is John F. Kennedy High School in Paterson, New Jersey
More Asian Students	More Asian Students	More Asian Students
–	–	More English Language Learner Students
–	–	More Hispanic Students
–	–	More White Students

We can now compare our best anomaly detection models built with classification-based techniques (Boosting, Voting Classifier, and Stochastic Gradient Descent) to our supervised implementation of the best-performing clustering-based technique (Gaussian Mixture). With supervised models, we prioritize the recall metric, as it indicates the proportion of dropouts successfully identified by the model. Referring back to Table 17 on page 47, we see the dropout percentages of the unsupervised Gaussian Mixture models are almost identical to the recall values displayed in Table 26, below. This reveals we can utilize these Gaussian Mixture models in both unsupervised and supervised learning situations without sacrificing performance.

Table 26

Best Clustering-Based Technique's Performance as a Supervised Anomaly Detection Tool

Best Clustering-Based Technique's Performance			
Supervised			
Metric	Column Set A (Gaussian Mixture)	Column Set B (Gaussian Mixture)	Column Set C (Gaussian Mixture)
Area Under the Curve	0.6281	0.6281	0.5622
Accuracy Score	0.6210	0.6210	0.5887
Precision	0.5588	0.5588	0.5625
Recall	0.6909	0.6909	0.3273

Although the performance metrics indicate Gaussian Mixture models effectively identify dropouts (specifically with Column Sets A and B), the three best classification techniques outperform all the distance-based and clustering-based approaches. The results in Table 27 show the classification-based models' numerical results are superior across all metrics. For example, the Boosting model built with Column Set A has a recall value of one; this is the highest possible value for this metric. It indicates the model is able to identify all dropouts in a dataset. Typically, this occurs when a model guesses one class for all individuals in the set; however, Column Set

Table 27

Best Classification-Based Technique's Performance as an Anomaly Detection Tool

Best Classification-Based Technique's Performance			
Metric	Column Set A (Boosting)	Column Set B (Voting Classifier)	Column Set C (Stochastic Gradient Descent)
Area Under the Curve	0.9583	0.6429	0.5595
Accuracy Score	0.9615	0.6538	0.5769
Precision	0.9333	0.6471	0.5789
Recall	1.0000	0.7857	0.7857

A's corresponding area under the curve and accuracy score values demonstrate this model's predictions differentiate between graduates and dropouts. These metrics consider the accuracy of predictions made across both groups. Their values reveal that approximately ninety-six percent of the model's twenty-six predictions are correct. If the classifier assumes all students would drop out, we would expect these values to be around forty-five percent; this value would correspond to our dataset's distribution of dropouts, which we discussed within Chapter Three's exploratory data analysis section. The confusion matrix in Figure 9 provides further confirmation that this classifier makes predictions with both groups.

Figure 9

Confusion Matrix for Column Set A's Boosting Model

	Predicted Graduates	Predicted Dropouts
Actual Graduates	11	1
Actual Dropouts	0	14

The Voting Classifier and Stochastic Gradient Descent models are also very accurate, as both recall values are 0.7857; when given a set of one hundred students who would drop out of college, these models would correctly identify approximately seventy-nine of the students. Comparing our results from the unsupervised and supervised

learning settings, it is clear that, numerically speaking, the classification-based models produce the most accurate anomaly detection tools.

Finally, we should look for any overlap among the important features of the supervised models. Changing the Gaussian Mixture models from unsupervised implementations to supervised ones reduces the number of student-centered characteristics from six (Table 19 on

Table 28

Best Clustering-Based Technique's Supervised Characteristics of Students Most Likely to Drop Out

Best Clustering-Based Technique's Supervised Findings: Student Characteristics		
Students With the Following Characteristics Are Likely to <u>DROP OUT</u>		
Column Set A (Gaussian Mixture)	Column Set B (Gaussian Mixture)	Column Set C (Gaussian Mixture)
African American/Black	African American/Black	–
Attends College Farther From Home	Attends College Farther From Home	–
Attends High School Farther From Home	Attends High School Farther From Home	Attends High School Farther From Home
Home Area Has Higher Assault Rates	Home Area Has Higher Assault Rates	–
Home Area Has Higher Crime Rates	Home Area Has Higher Crime Rates	–
Home Area Has Higher Murder Rates	Home Area Has Higher Murder Rates	–
Home Area Has Higher Personal Crime Rates	Home Area Has Higher Personal Crime Rates	–
Home Area Has Higher Rape Rates	Home Area Has Higher Rape Rates	–
Home Area Has Higher Robbery Rates	Home Area Has Higher Robbery Rates	–
Home Area Has Higher Vehicle Theft Rates	Home Area Has Higher Vehicle Theft Rates	–
Lives in a High Assault Area	Lives in a High Assault Area	–
Lives in a High Crime Area	Lives in a High Crime Area	–
Lives in a High Property Crime Area	Lives in a High Property Crime Area	–
Lives in a Larger Area	Lives in a Larger Area	–
–	–	More Absences

page 50) to one (Table 28, above); both approaches agree that students who live farther away from their high schools are at an increased risk for dropping out of college. Like their unsupervised counterparts, the supervised Gaussian Mixture models identify no patterns pertaining to the student characteristics of graduates (Table 29). As before, only Column Set C identifies “meaningful” associations.

Table 29

Best Clustering-Based Technique’s Supervised Characteristics of Students Most Likely to Graduate

Best Clustering-Based Technique’s Supervised Findings: Student Characteristics		
Students With the Following Characteristics Are Likely to <u>GRADUATE</u>		
Column Set A (Gaussian Mixture)	Column Set B (Gaussian Mixture)	Column Set C (Gaussian Mixture)
–	–	Attends College Farther From Home
–	–	Home Area Has Higher Murder Rates
–	–	Home Area Has Higher Vehicle Theft Rates
–	–	Takes at Least One Honors-Level Course

In order to compare these results to those of the classification-based techniques, we must highlight an important distinction between the two. For many classification models, the feature importance metrics do not distinguish between the two classes. Instead, an importance value is assigned to each predictor; higher values correspond to more important features. Hence, we cannot separate classification findings into dropout and graduate tables. In the context of this project, we conclude any factors that appear in all three column sets (highlighted in orange) have the most influence on if a student will drop out of college. However, we cannot determine whether that influence is positive or negative. For example, Table 30 a identifies a student’s distance between home and college, grade point average (GPA), and number of absences as the

Table 30

Best Classification-Based Technique's Student Characteristics Most Associated With Graduation Status

Best Classification-Based Technique's Findings: Student Characteristics		
The Following Student-Related Characteristics Are the Strongest Indicators of Graduation Status		
Column Set A (Boosting)	Column Set B (Voting Classifier)	Column Set C (Stochastic Gradient Descent)
–	–	African American/Black
–	–	Age at High School Graduation
–	Asian	–
Attending Bloomfield College	–	Attending Bloomfield College
–	–	Attending Kean University
–	–	Attending Marymount Manhattan College
–	–	Attending Montclair State University
–	–	Attending Passaic County Cmty. College
Attending Ramapo College of N.J.	–	Attending Ramapo College of N.J.
–	–	Attending Rutgers University
–	–	Attending The College of N.J.
Distance Between Home & College	Distance Between Home & College	Distance Between Home & College
Distance Between Home & High School	Distance Between Home & High School	–
Grade Point Average (GPA)	Grade Point Average (GPA)	Grade Point Average (GPA)
Hispanic	–	–
Home Area's Larceny Rates	–	–
–	Home Area's Rape Rates	–
–	–	Living in the 07504 ZIP Code
Multinational	Multinational	–
Number of Absences	Number of Absences	Number of Absences
Obtaining a Bachelor's Degree	–	–
Obtaining a STEM Degree	–	–
Obtaining an Associate's Degree	–	–
Participation in Community Service	–	–
SAT Math Score	–	–
SAT Writing Score	–	–
–	–	Takes at Least One Honors-Level Course

strongest indicators of graduation status. Unlike the distance-based and clustering-based techniques, we cannot use feature importance to determine if a farther distance between home and school is more common among dropouts or graduates. However, the models produced with Column Sets B and C provide coefficients for each variable. As such, we can use a coefficient's sign to determine if the variable is positively or negatively correlated with dropping out. In Column Set B, the coefficient of `Distance_to_PS` is approximately 286,115. Since this coefficient is positive, we know it is positively correlated with `PS_Graduated`. Therefore, this classification model indicates that students who live further away from home are more likely to drop out of college.

The classification-based techniques suggest no high school characteristics influence a student's graduation status. In other words, the individual's outcome is unaffected by his or her high school environment. From Table 31, we see only the model built with Column Set A identifies "meaningful" associations.

Table 31

Best Classification-Based Technique's High School Characteristics Most Associated With Graduation Status

Best Classification-Based Technique's Findings: High School Characteristics		
The Following School-Related Characteristics Are the Strongest Indicators of Graduation Status		
Column Set A (Boosting)	Column Set B (Voting Classifier)	Column Set C (Stochastic Gradient Descent)
Area's Vehicle Theft Rates	–	–
Percentage of Female Students	–	–
Percentage of Hispanic Students	–	–
Percentage of White Students	–	–

If we refer to Table 32, we find the supervised Gaussian Mixture models indicate schools in areas with high burglary rates and schools with higher quantities of Asian students often educate eventual college graduates. The latter association aligns with the findings of the distance-based and unsupervised clustering-based techniques. In Table 33, we see which characteristics the supervised Gaussian Mixture models identify as most frequently associated with the high schools of postsecondary dropouts. Surprisingly, only one of these five characteristics overlaps with the eight characteristics deemed important with unsupervised learning (Table 21 on page 53). Both learning styles assert schools with a large population of

Table 32

Best Clustering-Based Technique's Supervised Characteristics of High Schools Most Likely to Create Graduates

Best Clustering-Based Technique's Supervised Findings: High School Characteristics		
High Schools With the Following Characteristics Are Likely to Create <u>GRADUATES</u>		
Column Set A (Gaussian Mixture)	Column Set B (Gaussian Mixture)	Column Set C (Gaussian Mixture)
–	–	Area Has Higher Assault Rates
Area Has Higher Burglary Rates	Area Has Higher Burglary Rates	Area Has Higher Burglary Rates
–	–	Area Has Higher Crime Rates
–	–	Area Has Higher Personal Crime Rates
–	–	Area Has Higher Murder Rates
–	–	Area Has Higher Rape Rates
–	–	Area Has Higher Robbery Rates
–	–	More African American/Black Students
More Asian Students	More Asian Students	More Asian Students
More Hispanic Students	More Hispanic Students	–
–	–	More Homeless Students
–	–	More Students in Foster Care

Table 33

Best Clustering-Based Technique's Supervised Characteristics of High Schools Most Likely to Create Dropouts

Best Clustering-Based Technique's Supervised Findings: High School Characteristics		
High Schools With the Following Characteristics Are Likely to Create <u>DROPOUTS</u>		
Column Set A (Gaussian Mixture)	Column Set B (Gaussian Mixture)	Column Set C (Gaussian Mixture)
Area Has Higher Assault Rates	Area Has Higher Assault Rates	–
Area Has Higher Crime Rates	Area Has Higher Crime Rates	–
Area Has Higher Larceny Rates	Area Has Higher Larceny Rates	–
Area Has Higher Murder Rates	Area Has Higher Murder Rates	–
Area Has Higher Personal Crime Rates	Area Has Higher Personal Crime Rates	–
Area Has Higher Property Crime Rates	Area Has Higher Property Crime Rates	–
Area Has Higher Rape Rates	Area Has Higher Rape Rates	–
Area Has Higher Robbery Rates	Area Has Higher Robbery Rates	–
Area Has Higher Vehicle Theft Rates	Area Has Higher Vehicle Theft Rates	–
In a High Assault Area	In a High Assault Area	–
In a Larger Area	In a Larger Area	In a Larger Area
In a More Populous Area	In a More Populous Area	In a More Populous Area
More African American/Black Students	More African American/Black Students	–
–	–	More American Indian/Alaskan Native Students
–	–	More English Language Learner Students
–	–	More Hispanic Students
–	–	More Native Hawaiian/Pacific Islander Students
More Students	–	More Students
More Students in Foster Care	More Students in Foster Care	–
More Students w/ Free Lunch	More Students w/ Free Lunch	More Students w/ Free Lunch
More Students w/ Reduced Lunch	More Students w/ Reduced Lunch	More Students w/ Reduced Lunch
More White Students	More White Students	More White Students
More Teachers	–	More Teachers

students receiving reduced-cost lunches are more likely to educate students who will drop out of college.

Ultimately, there are no overlapping findings among the six supervised learning characteristic tables. Although these models achieve high recall values, they rely on different metrics to attain those results. Since our objective is to find a model that identifies which students are most likely to drop out of school before graduating college, we can consider the classification-based approach to be the best, as it is the one that attained the highest numerical performance. In the next chapter, we will explore how the characteristic findings of all three techniques can also be used to address college dropout rates.

We should now refit our best classification-based models to the union (i.e., the combination) of our training and validation datasets. Then, we can assess the performances of these three final models on the testing set. In Chapter Three, we noted the best performing classification-based model built with Column Set C used a scaled version of the dataset. Thus, we refit and assessed Model #3 using the scaled data. The results are presented in Table 34. The confusion matrices of the three models are displayed in Figures 10, 11, and 12.

Table 34

Performances of the Final Models on the Testing Set

Final Models' Performances on the Testing Set			
Metric	#1 Column Set A (Boosting)	#2 Column Set B (Voting Classifier)	#3 Column Set C (Stochastic Gradient Descent)
Area Under the Curve	1.0000	0.7500	0.3333
Accuracy Score	1.0000	0.7500	0.3333
Precision	1.0000	0.6667	0.3333
Recall	1.0000	1.0000	0.3333

Figure 10*Confusion Matrix for Final Model #1: Testing Set*

	Predicted Graduates	Predicted Dropouts
Actual Graduates	6	0
Actual Dropouts	0	6

Figure 11*Confusion Matrix for Final Model #2: Testing Set*

	Predicted Graduates	Predicted Dropouts
Actual Graduates	3	3
Actual Dropouts	0	6

Figure 12*Confusion Matrix for Final Model #3: Testing Set*

	Predicted Graduates	Predicted Dropouts
Actual Graduates	2	4
Actual Dropouts	4	2

Discussion

In Chapter Three, we explained how we selected the three column sets and highlighted the differences between them. We can now combine this information with the final models' numerical results to determine the optimal dataset. This dataset must be suitable, duplicatable, and effective. We will deem the model corresponding to that dataset the official deliverable of this thesis.

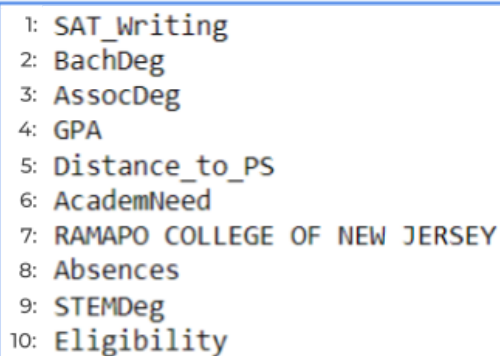
In Table 34, we see Model #1 attains perfect scores for all the metrics. Though this may seem impressive, predictive models that obtain perfect results often indicate an underlying association within the dataset and require further investigation. To address this possibility, we can compare the important features of Model #1 with those of one of the other models. As a reminder, we built all three models using the exact same parameters. The differing performance rates are a result of including or omitting information while modeling. Since the datasets used for Model #1 and Model #2 are almost identical (as shown in Appendices D and E), they are the easiest two models to compare. In Chapter Three, we explained how we created Column Set B by removing five of Column Set A's predictors, as we suspected these features were redundant with the information provided by other columns. For this reason, we also omitted these five variables from Column Set C. These features describe:

- 1) If the individual obtained an Associate's Degree
- 2) If the individual obtained a Bachelor's Degree
- 3) If the individual obtained a STEM Degree²
- 4) The total number of students at the individual's high school
- 5) The total number of teachers at the individual's high school

If any of these predictors appear in Model #1's list of most important features, this would indicate they are contributing to the discrepancy in model performance. Figure 13 displays this model's ten most important features. This list includes three of the five aforementioned variables, specifically: BachDeg, AssocDeg, and STEMDeg. Upon closer

Figure 13

The Ten Most Important Features in Model #1



```

1: SAT_Writing
2: BachDeg
3: AssocDeg
4: GPA
5: Distance_to_PS
6: AcademNeed
7: RAMAPO COLLEGE OF NEW JERSEY
8: Absences
9: STEMDeg
10: Eligibility

```

² A degree obtained in one of the following disciplines: Science, Technology, Engineering, or Mathematics

inspection of the summaries of these three predictors, we see one of their possible values indicates whether the student dropped out of college (see Appendix C). Initially, we sought to include these predictors to discover if there were any associations between the type of degree pursued and graduation status. Instead, Model #1 uses these features to extract if the student was a graduate or a dropout. Essentially, this model inadvertently contains four variables acting as the target. In practice, schools and administrators would not have this information during implementation. If we already know the student dropped out of college, it is too late to use the model as an early detection tool. Hence, Model #1 is impractical and cannot be considered the best model.

From a numeric standpoint, Table 34 (on page 64) suggests Model #2 is far superior to Model #3. However, since the testing set contains only twelve students, these metrics may not provide a fair assessment of the models. As such, we should also compare their characteristic findings to conduct a holistic assessment of the two early detection tools. Although Model #1 prioritizes impractical data while forming its predictions, it also references several other predictors. Since this model achieves such high accuracy, it is reasonable to deduce that whichever model that has the most overlapping important features with Model #1 will be the most successful. Table 35 displays each model's twenty most influential features list from most to least important. The highlighted cells indicate features in Models #1 and #2 (yellow), Models #1 and #3 (green), or all three models (orange) overlap. Model #2 has eight highlighted cells, while Model #3 only has four. Therefore, Model #2 numerically and characteristically outperforms Model #3. As such, our research and results support the denotation of Model #2 as our optimal and terminal model.

Table 35

The Twenty Most Important Features of Each Final Model (In Order of Importance)

The Most Important Features of Each Final Model		
#1 Column Set A (Boosting)	#2 Column Set B (Voting Classifier)	#3 Column Set C (Stochastic Gradient Descent)
SAT Writing Score	Size of Student's Area	Grade Point Average (GPA)
Obtaining a Bachelor's Degree	Size of School's Area	Taking at Least One Honors-Level Course
Obtaining an Associate's Degree	Population in School's Area	Attending Fortis Institute
Grade Point Average (GPA)	Population in Student's Area	School Is in the 07501 ZIP Code
Distance Between Home & College	Age at High School Graduation	Attending the College of N.J.
Foremost Academic Need	Distance Between Home & College	SAT Writing Score
Attending Ramapo College of N.J.	Number of Free Lunch Students	Attending Marymount Manhattan College
Number of Absences	SAT Math Score	Age at High School Graduation
Obtaining a STEM Degree	SAT Reading Score	Living in the 07112 ZIP Code
Reason for Eligibility in Upward Bound	SAT Writing Score	African American/Black
Home Area's Burglary Rates	Home Area's Robbery Rates	Attending Eastside High School
School Area's Vehicle Theft Rates	School Area's Robbery Rates	Attending Fairleigh Dickinson University
Age at High School Graduation	Number of Reduced Lunch Students	Number of Absences
Participation in Community Service	School Area's Burglary Rates	Living in the 07522 ZIP Code
Percentage of Black Students	Home Area's Murder Rates	Living in the 07108 ZIP Code
School Area's Burglary Rates	School Area's Personal Crime Rates	Attending Essex County College
Size of School's Area	School Area's Vehicle Theft Rates	Attending Daytona State College
Taking at Least One AP-Level Course	Percentage of White Students	School is Title I School-Wide
SAT Reading Score	Distance Between Home & High School	Attending Seton Hall University
Percentage of Female Students	Number of Absences	Attending Montclair State University

Note. Orange cells indicate all three models overlap. Yellow cells indicate Model #2 overlaps with Model #1. Green cells indicate Model #3 overlaps with Model #1.

As a final evaluation of our model, we can apply our anomaly detection tool to the reserved model deployment data. This allows us to simulate the efficacy of our model if we immediately implemented it within a school district. The results in Table 36 reveal the model identifies approximately sixty-nine percent of the dropouts. If we refer to the confusion matrix in Figure 14, we see this equates to missing only four students.

Table 36

Performance of the Terminal Model on the Model Deployment Set

Anomaly Detection Tool's Performance on the Model Deployment Dataset	
Metric	Terminal Model
Area Under the Curve	0.5128
Accuracy Score	0.5455
Precision	0.6000
Recall	0.6923

Figure 14

Confusion Matrix for the Best Model on Model Deployment

	Predicted Graduates	Predicted Dropouts
Actual Graduates	3	6
Actual Dropouts	4	9

We can also review the models' textual results to discover each technique's general description of a college dropout. While designing our research, we wanted to monitor each model's characteristic findings to determine which academic, institutional, personal, and social demographic information are most associated with dropping out of college. If we compile the characteristic findings of each anomaly detection technique, we obtain the comprehensive lists displayed in Table 37.

Across the three techniques, the models suggest there are twenty-eight features that influence a student's decision to graduate from or drop out of college. This indicates it would be sufficient to narrow our analysis from the initial one-hundred seventy-three variables to these twenty-eight. Only one of these features concerns academic performance: grade point average. Meanwhile, twenty-five of the identified predictors pertain to a student's environment, both at home and school. These findings support the results of the psychological studies referenced in Chapter Two, which indicated economic and environmental factors strongly influence a student's likelihood of dropping out of school (Muharrem, et al., 2020; Zorbaz & Özer, 2020).

The table also reveals there are two attributes all three techniques deem important: the number of times a student is absent throughout one year of high school and the distance between a student's home and college. Intuitively, it is easy to understand how these two variables are associated with a student's college graduation status. If a student regularly misses class in high school (a very structured and monitored environment), this habit would only be exacerbated in the college setting (a more unstructured and unsupervised setting). Similarly, students who attend college farther away from home are more distanced from those that could keep them on track and guide them toward success. With no support system, these students are more likely to make counterproductive decisions, causing them to drop out of school.

Table 37

A Comprehensive List of Which Factors Each Technique Deems Most Associated With Graduation Status

All Characteristic Findings of Each Technique		
Distance-Based	Clustering-Based	Classification-Based
African American/Black	–	–
Distance Between Home & College *	Distance Between Home & College *	Distance Between Home & College *
Distance Between Home & High School *	Distance Between Home & High School *	–
–	–	Grade Point Average (GPA) ●
Home Area's Burglary Rates *	–	–
–	Home Area's Murder Rates *	–
Home Area's Robbery Rates *	–	–
Home Area's Vehicle Theft Rates *	Home Area's Vehicle Theft Rates *	–
Number of Absences	Number of Absences	Number of Absences
# of African American/Black Students *	# of African American/Black Students *	–
# of Asian Students *	# of Asian Students *	–
# of Hispanic Students *	–	–
# of Homeless Students *	–	–
# of Students in Foster Care *	# of Students in Foster Care *	–
–	# of Students w/ Free Lunch *	–
–	# of Students w/ Reduced Lunch *	–
–	# of White Students *	–
Population in Home's Area *	–	–
–	Population in School's Area *	–
–	School Area's Assault Rates *	–
–	School Area's Burglary Rates *	–
School Area's Murder Rates *	School Area's Murder Rates *	–
School Area's Personal Crime Rates *	School Area's Personal Crime Rates *	–
School Area's Rape Rates *	–	–
School Area's Robbery Rates *	School Area's Robbery Rates *	–
School Area's Vehicle Theft Rates *	School Area's Vehicle Theft Rates *	–
Size of Home's Area *	Size of Home's Area *	–
Size of School's Area *	Size of School's Area *	–

Note. ♣ Denotes an environmental factor. ● Denotes an academic performance factor.

By comparing the numeric results and characteristic findings of each technique, we found the classification-based approaches most effectively identify eventual college dropouts. We assessed the predictors included in and the model performances of each column set and determined Column Set B is most suitable, duplicatable, and effective. Therefore, we denoted Final Model #2 as the optimal model. When used as an early detection tool, the model deployment data suggests the model will successfully identify approximately sixty-nine percent of the college dropouts within a dataset. All the anomaly detection techniques reported various student-centered and secondary school-centered characteristics that correlate with a student's graduation status. The three approaches all indicated the number of times a student is absent and the distance between a student's home and college are strongly associated with dropping out. The classification-based models' findings were limited in number and variety, while the distance-based and clustering-based models' findings were more well-rounded. In the next chapter, we will discuss how we could utilize this information to further improve our early detection tool.

Chapter Five: Conclusions

In Chapter One, we presented an overview of the increasing college dropout rates and shared our desire to create an effective early detection tool to help address this issue. In Chapter Two, we explored educators' current methods for identifying and supporting potential dropouts. We also highlighted various psychological studies, which found dropping out is most frequently related to economical and environmental factors. We explained how we used this information to form our research question. In an effort to provide an alternative to using test scores and academic performance to identify high-risk students, we combined a student's academic information with the empirical evidence of numerous environmental and economical factors to build a more holistic approach. In Chapter Three, we introduced the datasets we collected to approach this problem and detailed how we compiled and prepared these datasets for statistical modeling. We outlined three popular anomaly detection techniques: distance-based, clustering-based, and classification-based. We then defined twelve statistical models that employ techniques and reported their performances within this project. In Chapter Four, we compared the best-performing models to assess the efficacy of the three techniques. After comparing the numerical results and characteristic findings, we determined the Voting Classifier model built with Column Set B is the best numerical model, though the distance-based and clustering-based techniques provide a more comprehensive description of a student most likely to drop out of college.

In this final chapter, we share how our results could be used to create an innovative early detection tool that positively contributes to the field of education. We will describe how our analysis surpasses the current view of and approach to dropout rates, highlighting underlying

trends among college dropouts. We will also provide suggestions for schools who wish to create early detection tools that are tailored to their specific needs and desired outcomes. Finally, we will outline various ways to expand upon our research in an attempt to further increase model efficacy.

Contributions

From our research, we found it is possible to build models that predict which students are most likely to drop out of college with high accuracy. Specifically, classification-based Voting Classifier models can identify the largest percentage of dropouts within a dataset. These models rely on quantitative academic information such as a student's grade point average and SAT scores. These features are frequently viewed as the biggest indicators of success. However, only relying on these pieces of information can result in a large amount of variability. As we discussed in Chapter Two, these metrics can be very subjective. Moreover, studies show these scores are not effective assessments of a student's academic abilities (Shulman, 2018, p. 9).

When we approached the problem from an unsupervised perspective, rather than a supervised one, our models revealed there are numerous atypical characteristics that impact an individual's college graduation status. Furthermore, these findings are far more interpretable, as they differentiate between traits specific to dropouts and to graduates. Meanwhile, with the classification approaches, some models do not specify if the metrics are positively or negatively associated with dropping out. This makes it difficult to address these issues, as it is unclear which situations are most likely to lead to dropping out. Surprisingly, the unsupervised approaches indicate academic performance is not associated with a student's graduation status. This suggests that, although lower academic performance may indicate a student is likely to drop out, it is not the source of the problem. Rather, these outcomes are a manifestation of the

environmental issues affecting the student. As such, the current methods for reducing dropout rates only address the symptoms of dropping out. However, when we merely try to ameliorate the symptoms, “the problem only gets worse and creates more problems” (Franklin, 2019). Assigning students to remediation and test preparation classes may help increase their academic scores, but it will not decrease their likelihood of dropping out. Instead, we must determine the root causes and provide students with the appropriate coping strategies and resources to minimize its effects.

To do this, we should incorporate the overarching characteristic findings of the best unsupervised models into our early detection process. In practice, these models would be implemented in situations where the final outcome is undetermined. The goal is to identify students who are most likely to drop out and use intervention methods to prevent them from doing so. If we only employ the optimal classification-based model proposed in Chapter Four, it is likely many students will fall through the cracks, as the model uses a very narrow set of characteristics. If we only rely on the characteristics indicated by the distance-based and clustering-based approaches, we may end up with an overwhelming number of students. It would be impossible to sufficiently support all those students. Combining the two will generate smaller groups of students, allowing for more effective and individualized interventions. According to recent reports, “colleges and universities are moving away from seeing applicants only as test scores and grade point averages...holistic admissions are becoming the norm” (Newton, 2021). When reviewing applications, college admission officers consider “factors that are both academic and non-academic, as well as objective and subjective” (Newton). Since this approach is used when admitting students into college, we should employ the same tactics when trying to prevent students from leaving prematurely.

Before implementing these practices within a school district, it is important for administrators to use these techniques to build their own models. We built the models within this project to detect the underlying patterns of the data it trained on, specifically Ramapo College's Upward Bound Math Science students and their respective high schools. As we mentioned in previous chapters, these students represent a very distinct group of individuals. Across the United States, each school has a unique population and school environment; a school needs access to data containing this information to create useful models. Thus, each school should compile their own values for all features used in Column Set B. However, it is likely that limiting this information to the twenty-eight features listed in Table 37 on page 71 would be sufficient. Schools could then train k -Nearest Neighbors and Gaussian Mixture models as well as several classification-based models described in Chapter Three. They could then implement the procedures outlined above to ensure the models are tailored to their schools' specific needs and desired outcomes. Because we explored both supervised and unsupervised approaches within one project, this research provides a clear way to implement these methods in supervised, unsupervised, and semi-supervised settings. In this final scenario, schools would train models using historical data that indicates the college graduation status for each alumni and then apply this model to the students currently enrolled in the high school.

Future Work

In addition to the implementation expansions detailed in the previous section, there are several research-based approaches to explore that could extend the education field's understanding of and best practices for addressing college dropout rates. Some of the most promising and reasonable extensions are described below.

Density-Based Anomaly Detection

As we mentioned in Chapter Three, numerous techniques can be used to conduct anomaly detection. Various density-based techniques would be a next logical step for this research, as they utilize similar ideas to those of the distance-based and clustering-based methods. With density-based anomaly detection, the models assess “the density of an object and that of its neighbors;” any instance with a density “relatively much lower than that of its neighbors” is considered an outlier (Chepenko, 2018). This technique could be used as an additional unsupervised approach and could possibly provide further insight on which factors are most associated with dropping out.

Assessing for Differences in Graduation Rates Across Several Years

In this project, our limited access to quality data necessitated compiling all the students from the 2008 to 2015 high school graduating classes into one dataset. We performed the analysis on all students within our modeling dataset, separating the students into modeling and model deployment data without considering their graduation cohorts. With a larger dataset, it would be interesting to model and analyze the dropout rates and findings with one graduation class at a time. This way, it would be possible to track the differences in graduation rates and trends over the years. This could lead to further underlying associations that were undetectable in such a small dataset.

Reassess Students at the Culmination of Each School Year

As mentioned in Chapter One, an effective early detection tool would be implemented at the culmination of each school year. In this scenario, high school administrators would employ the models and their characteristic findings in May or June to identify which students are most at risk for dropping out in college. Throughout the following school year, the administration would

provide additional support to those students in an effort to decrease that risk. In May or June, the administrators would repeat this process. If the same students reappeared as high risk students, this would indicate they either misidentified the student's most prominent issue or implemented ineffective response strategies. This would provide model users with frequent feedback, which would help improve model implementation within the district. In this project, we only used the academic information from the students' senior year of high school. We could not implement the aforementioned sequential approach, as it would inadvertently assess the effectiveness of the Upward Bound program, rather than the intervention methods themselves.

Grid Search

In Chapter Three, we outlined how we manually modified the hyperparameters of the twelve model types throughout this project. In the future, it would be far more efficient to utilize a tool such as Scikit-Learn's `GridSearchCV`. This tool uses "cross-validation to evaluate all the possible combinations of hyperparameter values" (Géron, 2019, p. 76). Consequently, it would likely increase the accuracy of our models, as it is able to compare all of the possibilities to find the best model. This tool would also make it much easier for school administrators to incorporate this research into their districts. Schools would only need to compile the necessary information to align with Column Set B; the actual implementation of the research would be completely automated.

Combining the Supervised and Unsupervised Models

Earlier in this chapter, we explained how incorporating the outputs of both the unsupervised and supervised learning models would produce the best and most actionable results. In order to simplify this process for school districts, it would be advantageous to create a pipeline that automizes this process. First, users would apply the unsupervised tools to their

dataset to obtain a list of the most relevant features. Then, users would implement the supervised models to identify specific students. Preferably, the highest-performing classification-based model would be integrated with the most accurate unsupervised Gaussian Mixture model, as this method proved to be the best unsupervised approach throughout the project. Again, optimizing as much of this process as possible would encourage more school administrators to incorporate this research into their districts, providing support to a larger number of students across the country.

Determining When to Raise a Warning for a Potential Dropout

In Chapter Four, we highlighted how all three anomaly detection techniques agreed a student's number of absences and how far a student lives from college are strongly associated with dropping out of college. However, it is unclear which specific values would indicate an issue. Before implementing these models, it is important to determine when administrators should shift students from "on their radar" to "on their list for immediate action". For example, we know that more absences increase a student's likelihood of dropping out. Still, we need to identify the number of absences that signal a student as at-risk. Establishing these thresholds would make it easier for districts to adopt these models within their schools.

Although numerous psychological studies (like those discussed in Chapter Two) found economic and environmental factors have the largest impact upon a student's ability to graduate college, the education field's methods for identifying students most at risk for dropping out remain unchanged. Society tends to prefer quantitative research "over qualitative research because it is more scientific, objective, fast, focused[,] and acceptable" (Form Plus Blog, 2021). Within this project, we discovered a viable way to identify which students are most likely to drop

out of college using models that support these psychological findings. Because we developed a formalized way to utilize twenty years of qualitative findings, educators and administrators may be more willing to acknowledge these reports and modify their approaches to the increasing college dropout rates. In doing so, the schools will benefit from increased graduation rates. More importantly, the students in need who continue to struggle through school, undetected, can be identified and effectively guided toward success.

References

- Aravindharavindh. (2021, September 10). Clustering-based approaches for outlier detection in data mining. *GeeksforGeeks*.
<https://www.geeksforgeeks.org/clustering-based-approaches-for-outlier-detection-in-data-mining/>
- Boyd, S. & Vandenberghe, L. (2018). Norm and distance. *Introduction to applied linear algebra: Vectors, matrices, and least squares*. Cambridge University Press.
<https://doi.org/10.1017/9781108583664>
- Chepenko, D. (2018, September 15). A density-based algorithm for outlier detection. *Towards Data Science*.
<https://towardsdatascience.com/density-based-algorithm-for-outlier-detection-8f278d2f7983>
- Crime in the United States: Overview*. (2021, June 29). ArcGIS. <https://arcg.is/15STCH0>
- Editorial. (2021, June 19). Six anomaly detection techniques – Pros and cons. *Roboticsbiz*.
<https://roboticsbiz.com/six-anomaly-detection-techniques-pros-and-cons/>
- Esri (2019). *Esri's crime indexes* [Data set]. City of Columbus Maps & Apps.
<https://columbus.hub.arcgis.com/datasets/columbus::zip-code/about>
- Espressius, D. (2021, November 21). Anomaly detection I - Distance-based methods. *Dev*.
https://dev.to/_aadidev/anomaly-detection-i-distance-based-methods-278g
- Federal Student Aid (2022). *2022-23 federal school code list of participating schools (February 2022)* [Data set]. United States Department of Education.

<https://fsapartners.ed.gov/knowledge-center/library/federal-school-code-lists/2022-02-02/2022-23-federal-school-code-list-participating-schools-february-2022>

Federal TRIO programs: Home page. (2022, April 26). U.S. Department of Education.

<https://www2.ed.gov/about/offices/list/ope/trio/index.html>

Form Plus Blog. (2021, December 4). 15 reasons to choose quantitative over qualitative research.

Formplus. <https://www.formpl.us/blog/quantitative-qualitative-research>

Franklin, D. (2019, July 6). *Address the root cause, not the symptoms.* DavidFranklin.

<https://davidfranklin.org/address-the-root-cause-not-the-symptoms/>

Fuller, J. (2017, December 20). Why employers must stop requiring college degrees for middle-skill jobs. *Forbes*.

<https://www.forbes.com/sites/hbsworkingknowledge/2017/12/20/why-employers-must-stop-requiring-college-degrees-for-middle-skill-jobs/?sh=12dcdd4a4950>

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.)(R. Roumeliotis & N. Tache, Eds.). O'Reilly Media.

Goyal, A. (2019, July 18). Detecting the onset of machine failure using anomaly detection techniques. *Towards Data Science*.

<https://towardsdatascience.com/detecting-the-onset-of-machine-failure-using-anomaly-detection-techniques-d2f7a11eb809>

Gumbao, M. G. (2019, June 3). Best clustering algorithms for anomaly detection. *Towards Data Science*.

<https://towardsdatascience.com/best-clustering-algorithms-for-anomaly-detection-d5b7412537c8>

- Hacking the thesis.* (n.d.). The Ohio State University. Retrieved February 16, 2022, from <https://u.osu.edu/hackingthethesis/managing-stuff/your-content/outline/>
- Hiller, M. (2019, January 18). SAT and ACT exams do not accurately measure academic ability. *The Schreiber Times*.
<https://theschreibertimes.com/2019/01/18/sat-and-act-exams-do-not-accurately-measure-academic-ability>
- IBM Cloud Education. (2020, August 25). Exploratory data analysis. *IBM*.
<https://www.ibm.com/cloud/learn/exploratory-data-analysis>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer. <https://doi.org/10.1007/978-1-4614-7138-7>
- Leonhardt, D. & Chinoy, S. (2019, May 23). The college dropout crisis. *The New York Times*. Retrieved from <https://nyti.ms/32mShlf>
- McFarland, J., Cui, J., and Stark, P. (2018, February 22). *Trends in high school dropout and completion rates in the United States: 2014* (NCES 2018-117). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from <http://nces.ed.gov/pubsearch>
- Mehrotra, K. G., Mohan, C. K., & Huang, H. (2017). Clustering-based anomaly detection approaches. In *Anomaly Detection Principles and Algorithms* (pp. 41-55). Springer.
https://doi.org/10.1007/978-3-319-67526-8_4
- Muharrem, K., Osman, Z., & Demirtas-Zorbaz S. (2020). Has the ship sailed? The causes and consequences of school dropout from an ecological viewpoint. *Social Psychology of Education: An International Journal*, 23(5), 1149-1171.
<https://doi.org/10.1007/s11218-020-09568-w>

National Center for Education Statistics (n.d.). *About us*. Common Core of Data. Retrieved April 27, 2022, from <https://nces.ed.gov/about/>

National Center for Education Statistics (2021a). *Newark Public School District data for the 2020-2021 school year* [Data set]. Common Core of Data. Retrieved February 24, 2022, from https://nces.ed.gov/ccd/schoolsearch/school_list.asp?Search=1&DistrictID=3411340

National Center for Education Statistics (2021b). *Passaic County Technical Institute Public School data for the 2020-2021 school year* [Data set]. Common Core of Data. Retrieved February 24, 2022, from https://nces.ed.gov/ccd/schoolsearch/school_list.asp?Search=1&DistrictID=3412630

National Center for Education Statistics (2021c). *Paterson Public School District data for the 2020-2021 school year* [Data set]. Common Core of Data. Retrieved February 24, 2022, from https://nces.ed.gov/ccd/schoolsearch/school_list.asp?Search=1&DistrictID=3412690

New Jersey Department of Education (n.d.). *NJ school performance report*. The State of New Jersey. Retrieved April 27, 2022, from <https://rc.doe.state.nj.us/>

New Jersey Department of Education (2020). *2019-2020 NJ school performance report* [Data set]. The State of New Jersey. <https://rc.doe.state.nj.us/download>

Newton, K. (2021, April 30). Holistic admissions: Your child is more than their grades and test scores. *Road2College*. <https://www.road2college.com/holistic-admissions-your-child-is-more-than-their-grades-and-test-scores/>

Office of Postsecondary Education/Federal TRIO Programs. (2021). Upward Bound and Upward Bound Math-Science annual performance report: Appendix.

<https://www2.ed.gov/programs/triomathsci/report.html>

Pramoditha, R. (2020, July 20). Hands-on k-means clustering. *Medium*.

<https://medium.com/mlearning-ai/k-means-clustering-with-scikit-learn-e2af706450e4>

Ramapo College Upward Bound Math Science Program (2008-2021). *Annual performance reports* [Unpublished raw data]. Dr. Sandra Suarez (Director). Retrieved February 8, 2022.

Ranjan, A. (2020, November 30). Hierarchical clustering (Agglomerative). *Medium*.

[https://medium.com/analytics-vidhya/hierarchical-clustering-agglomerative-f6906d44098](https://medium.com/analytics-vidhya/hierarchical-clustering-agglomerative-f6906d440981)

[1](https://medium.com/analytics-vidhya/hierarchical-clustering-agglomerative-f6906d440981)

Rumberger, R. W., & Thomas, S. L. (2000). The distribution of dropout and turnover rates among urban and suburban high schools. *Sociology of Education*, 73(1), 39-67.

<https://doi.org/10.2307/2673198>

Shulman, J. (2018). The data we need for holistic admissions. *Change: The magazine of higher learning*, 50(5), 8–15. <https://doi.org/10.1080/00091383.2018.1510256>

Upward Bound program: Performance. (2021, October 29). U.S. Department of Education.

<https://www2.ed.gov/programs/trioupbound/performance.html>

Upward Bound program: Purpose. (2021, December 17). U.S. Department of Education.

<https://www2.ed.gov/programs/trioupbound/index.html>

Zorbaz, O., & Özer, A. (2020). Do student characteristics affecting school dropout risk differ from one school to another? *Egitim Ve Bilim*, 45(202), 191-210.

<https://doi.org/10.15390/EB.2020.8266>

Zuora (n.d.). *United States standard state codes* [Data set]. Knowledge Center.

https://knowledgecenter.zuora.com/BB_Introducing_Z_Business/D_Country%2C_State%2C_and_Province_Codes/B_State_Names_and_2-Digit_Codes#State_Names_and_Codes

Appendix A

Variables in the Final Dataset					
1	Gender	S	88	ALTIERUS CAREER EDUCATION	I
2	Hispanic	S	89	BERKELEY COLLEGE	I
3	Asian	S	90	BLOOMFIELD COLLEGE	I
4	Black	S	91	CENTRAL CONNECTICUT STATE UNIV	I
5	White	S	92	COLLEGE OF NEW JERSEY (THE)	I
6	Hawaiian	S	93	COMMUNITY COLLEGE OF RI - WARWICK	I
7	Multinational	S	94	DAYTONA STATE COLLEGE	I
8	LimitedEnglish	A	95	DEVRY UNIVERSITY	I
9	Eligibility	S	96	DREW UNIVERSITY	I
10	AcademNeed	A	97	ENTERPRISE STATE COMMUNITY COLLEGE	I
11	Grade_EnteredUB	A	98	ESSEX COUNTY COLLEGE	I
12	Participation	A	99	FAIRLEIGH DICKINSON UNIV-TEANECK	I
13	GPA	A	100	FORTIS INSTITUTE	I
14	HSGrad_Age	S	101	FRANCIS MARION UNIVERSITY	I
15	AcademAch_ELA	A	102	KEAN UNIVERSITY	I
16	AcademAch_Math	A	103	LINCOLN UNIVERSITY	I
17	Employed	S	104	MARYMOUNT MANHATTAN COLLEGE	I
18	CulturalAct	P	105	MIDDLESEX COUNTY COLLEGE	I
19	CommServ	P	106	MONTCLAIR STATE UNIVERSITY	I
20	Std_07011	P	107	NEW JERSEY INST OF TECHNOLOGY	I
21	Std_07026	P	108	NORTH CAROLINA CENTRAL UNIVERSITY	I
22	Std_07103	P	109	OAKWOOD UNIVERSITY	I
23	Std_07104	P	110	PASSAIC COUNTY COMMUNITY COLLEGE	I
24	Std_07105	P	111	RAMAPO COLLEGE OF NEW JERSEY	I
25	Std_07106	P	112	RIDER UNIVERSITY	I
26	Std_07107	P	113	ROWAN UNIVERSITY	I
27	Std_07108	P	114	RUTGERS, THE STATE UNIVERSITY OF NJ	I

28	Std_07112	P	115	SETON HALL UNIVERSITY	I
29	Std_07114	P	116	STANFORD UNIVERSITY	I
30	Std_07501	P	117	STEVENS INSTITUTE OF TECHNOLOGY	I
31	Std_07502	P	118	STOCKTON UNIVERSITY	I
32	Std_07503	P	119	SYRACUSE UNIVERSITY	I
33	Std_07504	P	120	TEMPLE UNIVERSITY	I
34	Std_07505	P	121	UNION COUNTY COLLEGE	I
35	Std_07506	P	122	UNIVERSITY OF NEW HAVEN	I
36	Std_07513	P	123	UNIVERSITY OF WISCONSIN-MADISON	I
37	Std_07514	P	124	VAUGHN COLLEGE OF AERONAUTICS AND TECHNO	I
38	Std_07522	P	125	VIRGINIA STATE UNIVERSITY	I
39	Std_07524	P	126	WILLIAM PATERSON UNIVERSITY	I
40	Absences	A	127	HS_CrimesIndex_2020	I
41	SAT_Reading	A	128	HS_PersonalCrimesIndex_2020	I
42	SAT_Math	A	129	HS_MurdersIndex_2020	I
43	SAT_Writing	A	130	HS_RapesIndex_2020	I
44	AP	A	131	HS_RobberiesIndex_2020	I
45	Honors	A	132	HS_AssaultsIndex_2020	I
46	AssocDeg	A	133	HS_PropertyCrimesIndex_2020	I
47	BachDeg	A	134	HS_BurglariesIndex_2020	I
48	STEMDeg	A	135	HS_LarceniesIndex_2020	I
49	Barringer High School	I	136	HS_VehicleTheftsIndex_2020	I
50	East Side High School	I	137	HS_AreaPopulation_2020	I
51	Eastside High School	I	138	HS_Area	I
52	John F. Kennedy High School	I	139	HS_TotalCrimesCateg	I
53	Malcolm X Shabazz High School	I	140	HS_PersonalCrimesCateg	I
54	Passaic County Technical Institute	I	141	HS_MurdersCateg	I
55	Rosa L. Parks School of Fine and Performing Arts	I	142	HS_RapesCateg	I
56	School of Earth and Space Science	I	143	HS_RobberiesCateg	I
57	School of Government and Public Administration	I	144	HS_AssaultsCateg	I
58	West Side High School	I	145	HS_PropertyCrimesCateg	I

59	HS_07103	I	146	HS_BurglariesCateg	I
60	HS_07104	I	147	HS_LarceniesCateg	I
61	HS_07105	I	148	HS_VehicleTheftsCateg	I
62	HS_07108	I	149	Student_CrimesIndex_2020	P
63	HS_07470	I	150	Student_PersonalCrimesIndex_2020	P
64	HS_07501	I	151	Student_MurdersIndex_2020	P
65	HS_07502	I	152	Student_RapesIndex_2020	P
66	HS_07505	I	153	Student_RobberiesIndex_2020	P
67	HS_07514	I	154	Student_AssaultsIndex_2020	P
68	Title1_School	I	155	Student_PropertyCrimesIndex_2020	P
69	Title1_SchoolWide	I	156	Student_BurglariesIndex_2020	P
70	NumStudents	I	157	Student_LarceniesIndex_2020	P
71	NumTeachers	I	158	Student_VehicleTheftsIndex_2020	P
72	StudentTeacherRatio	I	159	Student_AreaPopulation_2020	P
73	NumFreeLunch	I	160	Student_Area	P
74	NumReducedLunch	I	161	Student_TotalCrimesCateg	P
75	%_Female	I	162	Student_PersonalCrimesCateg	P
76	%_Male	I	163	Student_MurdersCateg	P
77	%_EconomDisadv	I	164	Student_RapesCateg	P
78	%_w/Disabilities	I	165	Student_RobberiesCateg	P
79	%_ELL	I	166	Student_AssaultsCateg	P
80	%_Homeless	I	167	Student_PropertyCrimesCateg	P
81	%_FosterCare	I	168	Student_BurglariesCateg	P
82	%_White	I	169	Student_LarceniesCateg	P
83	%_Hispanic	I	170	Student_VehicleTheftsCateg	P
84	%_Black	I	171	Distance_to_HS	P
85	%_Asian	I	172	Distance_to_PS	P
86	%_Hawaiian	I	173	PS_Graduated	A
87	%_AmericanIndian	I			

Appendix A. A comprehensive list of the 173 variables in the final dataset. Red columns are directly from Ramapo College Upward Bound Math Science data. Orange columns are directly from National Center for Education Statistics data. Yellow columns are directly from New Jersey Department of Education data. Green columns are directly from Esri data. The blue letters indicate if each variable provides academic (A), personal (P), institutional (I), or social demographic (S) information.

Appendix B

Using Linear Algebra to Calculate the Geographic Distances

In linear algebra, we can compute the distance between two locations using their three-dimensional coordinates. Typically, “a location on the earth’s surface is...given by its latitude θ and its longitude λ , which correspond to angular distance from the equator and prime meridian, respectively” (Boyd & Vandenberghe, 2018, p. 66). The corresponding coordinates are given by the 3-vector shown in Equation (2), where R is 6,367.5 (the approximate radius of the earth, in kilometers) and θ and λ are in radians.

$$\begin{bmatrix} R \sin \lambda \cos \theta \\ R \cos \lambda \cos \theta \\ R \sin \theta \end{bmatrix} \quad (2)$$

In the context of this research, the most appropriate distance metric is the distance along the surface of the earth. We compute this with Equation (3), the product of the earth’s radius and the angle between the two vectors.

$$R\Delta(a, b) = R \arccos\left(\frac{a^T b}{\|a\| \|b\|}\right) \quad (3)$$

Appendix C

Description of Each Variable		
1	Gender	Student's Gender 1: Female 0: Male -1: Unknown
2	Hispanic	Student is identified as Hispanic/Latino 1: Yes 0: No
3	Asian	Student is identified as Asian 1: Yes 0: No
4	Black	Student is identified as Black or African American 1: Yes 0: No
5	White	Student is identified as White 1: Yes 0: No
6	Hawaiian	Student is identified as Native Hawaiian or Other Pacific Islander 1: Yes 0: No
7	Multinational	Student is identified as one or more of the aforementioned ethnicities and races 1: Yes 0: No
8	LimitedEnglish	Student has limited English proficiency at time of initial selection into Upward Bound 1: Yes 0: No
9	Eligibility	Reason student is eligible for Upward Bound 1: Low Income and First Generation 2: Low Income Only 3: First Generation Only 4: At Risk for Academic Failure Only 5: Low Income and At High Risk for Academic Failure 6: First Generation and At High Risk for Academic Failure 7: Low Income, First Generation, and At High Risk for Academic Failure -1: Unknown
10	AcademNeed	Student's other academic needs at time of initial selection into Upward Bound 1: Low GPA 2: Low Achievement Test Scores 3: Low Educational Aspirations

		4: Low GPA and Low Educational Aspirations 5: Low GPA and Low Achievement Test Scores 6: Low Achievement Test Scores and Low Educational Aspirations 7: Lack of Opportunity/Support 8: Lack of Career Goals 9: Limited English Proficiency 10: Lack of Confidence/Self Esteem/Social Skills 11: Predominately Low Income Community 12: Rural Isolation 13: Interest in Careers in Math and Science 14: Other 15: Diagnosed Learning Disability -1: Unknown
11	Grade_EnteredUB	Student's grade level at first Upward Bound service 8: Rising 9th grader (summer between 8th and 9th grade) 9: 9th grader 10: Rising 10th grader (summer between 9th and 10th grade) or 10th grader 11: Rising 11th grader (summer between 10th and 11th grade) or 11th grader 12: Rising 12th grader (summer between 11th and 12th grade) 13: 12th grader 14: Other 15: In 5th year of high school
12	Participation	Student's participation level in Upward Bound during the reporting year 1: Academic Year And Summer Components 2: Academic Year And Summer Bridge 3: Academic Year Only 4: Summer Component Only 5: Summer Bridge Only 6: Prior-Year Participant -1: Unknown
13	GPA	Student's high school cumulative grade point average at the end of the reporting year
14	HSGrad_Age	Student's age, in days, upon graduating high school
15	AcademAch_ELA	Student achieved the proficient level on state high school reading/language arts assessments 1: Yes 0: No 9: Not Applicable -1: Unknown
16	AcademAch_Math	Student achieved the proficient level on state high school math assessments 1: Yes 0: No 9: Not Applicable -1: Unknown
17	Employed	Student has a job during the reporting year 1: Yes 0: No 9: Not Applicable -1: Unknown

18	CulturalAct	Student participates in cultural activities offered by Upward Bound during the reporting year 1: Yes 0: No 9: Not Applicable -1: Unknown
19	CommServ	Student participates in community service activities offered by Upward Bound during the reporting year 1: Yes 0: No 9: Not Applicable -1: Unknown
20	Std_07011	Student lives in the 07011 ZIP Code
21	Std_07026	Student lives in the 07026 ZIP Code
22	Std_07103	Student lives in the 07103 ZIP Code
23	Std_07104	Student lives in the 07104 ZIP Code
24	Std_07105	Student lives in the 07105 ZIP Code
25	Std_07106	Student lives in the 07106 ZIP Code
26	Std_07107	Student lives in the 07107 ZIP Code
27	Std_07108	Student lives in the 07108 ZIP Code
28	Std_07112	Student lives in the 07112 ZIP Code
29	Std_07114	Student lives in the 07114 ZIP Code
30	Std_07501	Student lives in the 07501 ZIP Code
31	Std_07502	Student lives in the 07502 ZIP Code
32	Std_07503	Student lives in the 07503 ZIP Code
33	Std_07504	Student lives in the 07504 ZIP Code
34	Std_07505	Student lives in the 07505 ZIP Code
35	Std_07506	Student lives in the 07506 ZIP Code
36	Std_07513	Student lives in the 07513 ZIP Code
37	Std_07514	Student lives in the 07514 ZIP Code
38	Std_07522	Student lives in the 07522 ZIP Code
39	Std_07524	Student lives in the 07524 ZIP Code
40	Absences	Student's total number of absences during senior year of high school
41	SAT_Reading	Student's highest score on the SAT Reading section
42	SAT_Math	Student's highest score on the SAT Math section
43	SAT_Writing	Student's highest score on the SAT Writing section

44	AP	1: Student took at least one AP course in high school 0: Student did not take an AP course in high school
45	Honors	1: Student took at least one honors course in high school 0: Student did not take an honors course in high school
46	AssocDeg	Associate's degree attained 1: Yes 2: No; Pursued Another Kind of Postsecondary Credential 3: Deferred Enrollment 5: Transferred to Four-Year Without Completing Associate 6: Currently Enrolled 7: Left Postsecondary Program without Completing It 9: Not Applicable -1: Unknown
47	BachDeg	Bachelor's degree attained 1: Yes 2: No; Pursued Another Kind of Postsecondary Credential 3: Deferred Enrollment 6: Currently Enrolled 7: Left Postsecondary Program without Completing It 9: Not Applicable -1: Unknown
48	STEMDeg	Postsecondary STEM degree 1: Yes; "Hard" Science 2: Yes; Psychology or Social Science 0: No 9: Not Applicable -1: Unknown
49	Barringer High School	Student attended Barringer High School in Newark, NJ
50	East Side High School	Student attended East Side High School in Newark, NJ
51	Eastside High School	Student attended Eastside High School in Paterson, NJ
52	John F. Kennedy High School	Student attended John F. Kennedy High School in Paterson, NJ
53	Malcolm X Shabazz High School	Student attended Malcolm X Shabazz High School in Newark, NJ
54	Passaic County Technical Institute	Student attended Passaic County Technical Institute in Wayne, NJ
55	Rosa L. Parks School of Fine and Performing Arts	Student attended Rosa L. Parks School of Fine and Performing Arts in Paterson, NJ
56	School of Earth and Space Science	Student attended School of Earth and Space Science in Paterson, NJ
57	School of Government and Public Administration	Student attended School of Government and Public Administration in Paterson, NJ
58	West Side High School	Student attended West Side High School in Newark, NJ
59	HS_07103	Student's high school is in the 07103 ZIP Code
60	HS_07104	Student's high school is in the 07104 ZIP Code

61	HS_07105	Student's high school is in the 07105 ZIP Code
62	HS_07108	Student's high school is in the 07108 ZIP Code
63	HS_07470	Student's high school is in the 07470 ZIP Code
64	HS_07501	Student's high school is in the 07501 ZIP Code
65	HS_07502	Student's high school is in the 07502 ZIP Code
66	HS_07505	Student's high school is in the 07505 ZIP Code
67	HS_07514	Student's high school is in the 07514 ZIP Code
68	Title1_School	Student's high school is a Title I school
69	Title1_SchoolWide	Student's high school has a Title I school-wide program
70	NumStudents	Total number of students in the student's high school
71	NumTeachers	Total number of teachers in the student's high school
72	StudentTeacherRatio	Student-teacher ratio in the student's high school
73	NumFreeLunch	Number of students at the student's high school receiving free lunch
74	NumReducedLunch	Number of students at the student's high school receiving reduced lunch
75	%_Female	Percentage of students who are female in the student's high school
76	%_Male	Percentage of students who are male in the student's high school
77	%_EconomDisadv	Percentage of students who are economically disadvantaged in the student's high school
78	%_w/Disabilities	Percentage of students with disabilities in the student's high school
79	%_ELL	Percentage of students who are English language learners in the student's high school
80	%_Homeless	Percentage of students who are homeless in the student's high school
81	%_FosterCare	Percentage of students who are in Foster Care in the student's high school
82	%_White	Percentage of students who are White in the student's high school
83	%_Hispanic	Percentage of students who are Hispanic in the student's high school
84	%_Black	Percentage of students who are Black or African American in the student's high school
85	%_Asian	Percentage of students who are Asian in the student's high school

86	%_Hawaiian	Percentage of students who are Native Hawaiian or Pacific Islander in the student's high school
87	%_AmericanIndian	Percentage of students who are American Indian or Alaska Native in the student's high school
88	ALTIERUS CAREER EDUCATION	Student attended Altierus Career College
89	BERKELEY COLLEGE	Student attended Berkeley College
90	BLOOMFIELD COLLEGE	Student attended Bloomfield College
91	CENTRAL CONNECTICUT STATE UNIV	Student attended Central Connecticut State University
92	COLLEGE OF NEW JERSEY (THE)	Student attended The College of New Jersey (TCNJ)
93	COMMUNITY COLLEGE OF RI - WARWICK	Student attended Community College of Rhode Island (Warwick)
94	DAYTONA STATE COLLEGE	Student attended Daytona State College
95	DEVRY UNIVERSITY	Student attended DeVry University
96	DREW UNIVERSITY	Student attended Drew University
97	ENTERPRISE STATE COMMUNITY COLLEGE	Student attended Enterprise State Community College
98	ESSEX COUNTY COLLEGE	Student attended Essex County College
99	FAIRLEIGH DICKINSON UNIV-TEANECK	Student attended Fairleigh Dickinson University (Teaneck)
100	FORTIS INSTITUTE	Student attended Fortis Institute
101	FRANCIS MARION UNIVERSITY	Student attended Francis Marion University
102	KEAN UNIVERSITY	Student attended Kean University
103	LINCOLN UNIVERSITY	Student attended Lincoln University
104	MARYMOUNT MANHATTAN COLLEGE	Student attended Marymount Manhattan College
105	MIDDLESEX COUNTY COLLEGE	Student attended Middlesex County College
106	MONTCLAIR STATE UNIVERSITY	Student attended Montclair State University
107	NEW JERSEY INST OF TECHNOLOGY	Student attended New Jersey Institute of Technology (NJIT)
108	NORTH CAROLINA CENTRAL UNIVERSITY	Student attended North Carolina Central University
109	OAKWOOD UNIVERSITY	Student attended Oakwood University
110	PASSAIC COUNTY COMMUNITY COLLEGE	Student attended Passaic County Community College
111	RAMAPO COLLEGE OF NEW JERSEY	Student attended Ramapo College of New Jersey
112	RIDER UNIVERSITY	Student attended Rider University
113	ROWAN UNIVERSITY	Student attended Rowan University
114	RUTGERS, THE STATE UNIVERSITY OF NJ	Student attended Rutgers, The State University of New Jersey
115	SETON HALL UNIVERSITY	Student attended Seton Hall University

116	STANFORD UNIVERSITY	Student attended Stanford University
117	STEVENS INSTITUTE OF TECHNOLOGY	Student attended Stevens Institute of Technology
118	STOCKTON UNIVERSITY	Student attended Stockton University
119	SYRACUSE UNIVERSITY	Student attended Syracuse University
120	TEMPLE UNIVERSITY	Student attended Temple University
121	UNION COUNTY COLLEGE	Student attended Union County College
122	UNIVERSITY OF NEW HAVEN	Student attended University of New Haven
123	UNIVERSITY OF WISCONSIN-MADISON	Student attended University of Wisconsin–Madison
124	VAUGHN COLLEGE OF AERONAUTICS AND TECHNO	Student attended Vaughn College of Aeronautic Engineering, Aviation, and Technology
125	VIRGINIA STATE UNIVERSITY	Student attended Virginia State University
126	WILLIAM PATERSON UNIVERSITY	Student attended William Paterson University
127	HS_CrimesIndex_2020	Total Crime Index in student's high school's area
128	HS_PersonalCrimesIndex_2020	Personal Crime Index in student's high school's area
129	HS_MurdersIndex_2020	Murder Index in student's high school's area
130	HS_RapesIndex_2020	Rape Index in student's high school's area
131	HS_RobberiesIndex_2020	Robbery Index in student's high school's area
132	HS_AssaultsIndex_2020	Assault Index in student's high school's area
133	HS_PropertyCrimesIndex_2020	Property Crime Index in student's high school's area
134	HS_BurglariesIndex_2020	Burglary Index in student's high school's area
135	HS_LarceniesIndex_2020	Larceny Index in student's high school's area
136	HS_VehicleTheftsIndex_2020	Vehicle Theft Index in student's high school's area
137	HS_AreaPopulation_2020	Total Population in student's high school's area
138	HS_Area	Size of the student's high school's area
139	HS_TotalCrimesCateg	Level of total crime in student's high school's area 1.0: Low 0.0: High
140	HS_PersonalCrimesCateg	Level of personal crime in student's high school's area 1.0: Low 0.0: High
141	HS_MurdersCateg	Level of murder crime in student's high school's area 1.0: Low 0.0: High
142	HS_RapesCateg	Level of rape crime in student's high school's area 1.0: Low 0.0: High

143	HS_RobberiesCateg	Level of robbery crime in student's high school's area 1.0: Low 0.0: High
144	HS_AssaultsCateg	Level of assault crime in student's high school's area 1.0: Low 0.0: High
145	HS_PropertyCrimesCateg	Level of property crime in student's high school's area 1.0: Low 0.0: High
146	HS_BurglariesCateg	Level of burglary crime in student's high school's area 1.0: Low 0.0: High
147	HS_LarceniesCateg	Level of larceny crime in student's high school's area 1.0: Low 0.0: High
148	HS_VehicleTheftsCateg	Level of vehicle theft crime in student's high school's area 1.0: Low 0.0: High
149	Student_CrimesIndex_2020	Total Crime Index in student's home area
150	Student_PersonalCrimesIndex_2020	Personal Crime Index in student's home area
151	Student_MurdersIndex_2020	Murder Index in student's home area
152	Student_RapesIndex_2020	Rape Index in student's home area
153	Student_RobberiesIndex_2020	Robbery Index in student's home area
154	Student_AssaultsIndex_2020	Assault Index in student's home area
155	Student_PropertyCrimesIndex_2020	Property Crime Index in student's home area
156	Student_BurglariesIndex_2020	Burglary Index in student's home area
157	Student_LarceniesIndex_2020	Larceny Index in student's home area
158	Student_VehicleTheftsIndex_2020	Vehicle Theft Index in student's home area
159	Student_AreaPopulation_2020	Total Population in student's home area
160	Student_Area	Size of the student's home area
161	Student_TotalCrimesCateg	Level of total crime in student's home area 1.0: Low 0.0: High
162	Student_PersonalCrimesCateg	Level of personal crime in student's home area 1.0: Low 0.0: High
163	Student_MurdersCateg	Level of murder crime in student's home area 1.0: Low 0.0: High

164	Student_RapesCateg	Level of rape crime in student's home area 1.0: Low 0.0: High
165	Student_RobberiesCateg	Level of robbery crime in student's home area 1.0: Low 0.0: High
166	Student_AssaultsCateg	Level of assault crime in student's home area 1.0: Low 0.0: High
167	Student_PropertyCrimesCateg	Level of property crime in student's home area 1.0: Low 0.0: High
168	Student_BurglariesCateg	Level of burglary crime in student's home area 1.0: Low 0.0: High
169	Student_LarceniesCateg	Level of larceny crime in student's home area 1.0: Low 0.0: High
170	Student_VehicleTheftsCateg	Level of vehicle theft crime in student's home area 1.0: Low 0.0: High
171	Distance_to_HS	Distance along the surface of the earth between the student's home and high school
172	Distance_to_PS	Distance along the surface of the earth between the student's home and college
173	PS_Graduated	Student graduated from college 1: Yes 0: No 9: Not Applicable -1: Unknown

Appendix C. A brief description of each of the 173 variables in the final dataset.

Appendix D

Variables in Column Set A			
1	Gender	87	BERKELEY COLLEGE
2	Hispanic	88	BLOOMFIELD COLLEGE
3	Asian	89	CENTRAL CONNECTICUT STATE UNIV
4	Black	90	COLLEGE OF NEW JERSEY (THE)
5	White	91	COMMUNITY COLLEGE OF RI - WARWICK
6	Hawaiian	92	DAYTONA STATE COLLEGE
7	Multinational	93	DEVRY UNIVERSITY
8	LimitedEnglish	94	DREW UNIVERSITY
9	Eligibility	95	ENTERPRISE STATE COMMUNITY COLLEGE
10	AcademNeed	96	ESSEX COUNTY COLLEGE
11	Grade_EnteredUB	97	FAIRLEIGH DICKINSON UNIV-TEANECK
12	Participation	98	FORTIS INSTITUTE
13	GPA	99	FRANCIS MARION UNIVERSITY
14	HSGrad_Age	100	KEAN UNIVERSITY
15	AcademAch_ELA	101	LINCOLN UNIVERSITY
16	AcademAch_Math	102	MARYMOUNT MANHATTAN COLLEGE
17	Employed	103	MIDDLESEX COUNTY COLLEGE
18	CulturalAct	104	MONTCLAIR STATE UNIVERSITY
19	CommServ	105	NEW JERSEY INST OF TECHNOLOGY
20	Std_07011	106	NORTH CAROLINA CENTRAL UNIVERSITY
21	Std_07026	107	OAKWOOD UNIVERSITY
22	Std_07103	108	PASSAIC COUNTY COMMUNITY COLLEGE
23	Std_07104	109	RAMAPO COLLEGE OF NEW JERSEY
24	Std_07105	110	RIDER UNIVERSITY
25	Std_07106	111	ROWAN UNIVERSITY
26	Std_07107	112	RUTGERS, THE STATE UNIVERSITY OF NJ
27	Std_07108	113	SETON HALL UNIVERSITY

28	Std_07112	114	STANFORD UNIVERSITY
29	Std_07114	115	STEVENS INSTITUTE OF TECHNOLOGY
30	Std_07501	116	STOCKTON UNIVERSITY
31	Std_07502	117	SYRACUSE UNIVERSITY
32	Std_07503	118	TEMPLE UNIVERSITY
33	Std_07504	119	UNION COUNTY COLLEGE
34	Std_07505	120	UNIVERSITY OF NEW HAVEN
35	Std_07506	121	UNIVERSITY OF WISCONSIN-MADISON
36	Std_07513	122	VAUGHN COLLEGE OF AERONAUTICS AND TECHNO
37	Std_07514	123	VIRGINIA STATE UNIVERSITY
38	Std_07522	124	WILLIAM PATERSON UNIVERSITY
39	Std_07524	125	HS_CrimesIndex_2020
40	Absences	126	HS_PersonalCrimesIndex_2020
41	SAT_Reading	127	HS_MurdersIndex_2020
42	SAT_Math	128	HS_RapesIndex_2020
43	SAT_Writing	129	HS_RobberiesIndex_2020
44	AP	130	HS_AssaultsIndex_2020
45	Honors	131	HS_PropertyCrimesIndex_2020
46	AssocDeg	132	HS_BurglariesIndex_2020
47	BachDeg	133	HS_LarceniesIndex_2020
48	STEMDeg	134	HS_VehicleTheftsIndex_2020
49	Barringer High School	135	HS_AreaPopulation_2020
50	East Side High School	136	HS_Area
51	Eastside High School	137	HS_TotalCrimesCateg
52	John F. Kennedy High School	138	HS_PersonalCrimesCateg
53	Malcolm X Shabazz High School	139	HS_MurdersCateg
54	Passaic County Technical Institute	140	HS_RapesCateg
55	Rosa L. Parks School of Fine and Performing Arts	141	HS_RobberiesCateg
56	School of Earth and Space Science	142	HS_AssaultsCateg
57	School of Government and Public Administration	143	HS_PropertyCrimesCateg
58	West Side High School	144	HS_BurglariesCateg

59	HS_07103	145	HS_LarceniesCateg
60	HS_07104	146	HS_VehicleTheftsCateg
61	HS_07105	147	Student_CrimesIndex_2020
62	HS_07108	148	Student_PersonalCrimesIndex_2020
63	HS_07470	149	Student_MurdersIndex_2020
64	HS_07501	150	Student_RapesIndex_2020
65	HS_07502	151	Student_RobberiesIndex_2020
66	HS_07505	152	Student_AssaultsIndex_2020
67	HS_07514	153	Student_PropertyCrimesIndex_2020
68	Title1_School	154	Student_BurglariesIndex_2020
69	Title1_SchoolWide	155	Student_LarceniesIndex_2020
70	NumStudents	156	Student_VehicleTheftsIndex_2020
71	NumTeachers	157	Student_AreaPopulation_2020
72	StudentTeacherRatio	158	Student_Area
73	NumFreeLunch	159	Student_TotalCrimesCateg
74	NumReducedLunch	160	Student_PersonalCrimesCateg
75	%_Female	161	Student_MurdersCateg
76	%_Male	162	Student_RapesCateg
77	%_EconomDisadv	163	Student_RobberiesCateg
78	%_w/Disabilities	164	Student_AssaultsCateg
79	%_ELL	165	Student_PropertyCrimesCateg
80	%_Homeless	166	Student_BurglariesCateg
81	%_FosterCare	167	Student_LarceniesCateg
82	%_White	168	Student_VehicleTheftsCateg
83	%_Hispanic	169	Distance_to_HS
84	%_Black	170	Distance_to_PS
85	%_Asian	171	PS_Graduated
86	ALTIERUS CAREER EDUCATION		

Appendix D. A comprehensive list of the 171 variables in Column Set A.

Appendix E

Variables in Column Set B			
1	Gender	84	CENTRAL CONNECTICUT STATE UNIV
2	Hispanic	85	COLLEGE OF NEW JERSEY (THE)
3	Asian	86	COMMUNITY COLLEGE OF RI - WARWICK
4	Black	87	DAYTONA STATE COLLEGE
5	White	88	DEVRY UNIVERSITY
6	Hawaiian	89	DREW UNIVERSITY
7	Multinational	90	ENTERPRISE STATE COMMUNITY COLLEGE
8	LimitedEnglish	91	ESSEX COUNTY COLLEGE
9	Eligibility	92	FAIRLEIGH DICKINSON UNIV-TEANECK
10	AcademNeed	93	FORTIS INSTITUTE
11	Grade_EnteredUB	94	FRANCIS MARION UNIVERSITY
12	Participation	95	KEAN UNIVERSITY
13	GPA	96	LINCOLN UNIVERSITY
14	HSGrad_Age	97	MARYMOUNT MANHATTAN COLLEGE
15	AcademAch_ELA	98	MIDDLESEX COUNTY COLLEGE
16	AcademAch_Math	99	MONTCLAIR STATE UNIVERSITY
17	Employed	100	NEW JERSEY INST OF TECHNOLOGY
18	CulturalAct	101	NORTH CAROLINA CENTRAL UNIVERSITY
19	CommServ	102	OAKWOOD UNIVERSITY
20	Std_07011	103	PASSAIC COUNTY COMMUNITY COLLEGE
21	Std_07026	104	RAMAPO COLLEGE OF NEW JERSEY
22	Std_07103	105	RIDER UNIVERSITY
23	Std_07104	106	ROWAN UNIVERSITY
24	Std_07105	107	RUTGERS, THE STATE UNIVERSITY OF NJ
25	Std_07106	108	SETON HALL UNIVERSITY
26	Std_07107	109	STANFORD UNIVERSITY
27	Std_07108	110	STEVENS INSTITUTE OF TECHNOLOGY

28	Std_07112	111	STOCKTON UNIVERSITY
29	Std_07114	112	SYRACUSE UNIVERSITY
30	Std_07501	113	TEMPLE UNIVERSITY
31	Std_07502	114	UNION COUNTY COLLEGE
32	Std_07503	115	UNIVERSITY OF NEW HAVEN
33	Std_07504	116	UNIVERSITY OF WISCONSIN-MADISON
34	Std_07505	117	VAUGHN COLLEGE OF AERONAUTICS AND TECHNO
35	Std_07506	118	VIRGINIA STATE UNIVERSITY
36	Std_07513	119	WILLIAM PATERSON UNIVERSITY
37	Std_07514	120	HS_CrimesIndex_2020
38	Std_07522	121	HS_PersonalCrimesIndex_2020
39	Std_07524	122	HS_MurdersIndex_2020
40	Absences	123	HS_RapesIndex_2020
41	SAT_Reading	124	HS_RobberiesIndex_2020
42	SAT_Math	125	HS_AssaultsIndex_2020
43	SAT_Writing	126	HS_PropertyCrimesIndex_2020
44	AP	127	HS_BurglariesIndex_2020
45	Honors	128	HS_LarceniesIndex_2020
46	Barringer High School	129	HS_VehicleTheftsIndex_2020
47	East Side High School	130	HS_AreaPopulation_2020
48	Eastside High School	131	HS_Area
49	John F. Kennedy High School	132	HS_TotalCrimesCateg
50	Malcolm X Shabazz High School	133	HS_PersonalCrimesCateg
51	Passaic County Technical Institute	134	HS_MurdersCateg
52	Rosa L. Parks School of Fine and Performing Arts	135	HS_RapesCateg
53	School of Earth and Space Science	136	HS_RobberiesCateg
54	School of Government and Public Administration	137	HS_AssaultsCateg
55	West Side High School	138	HS_PropertyCrimesCateg
56	HS_07103	139	HS_BurglariesCateg
57	HS_07104	140	HS_LarceniesCateg
58	HS_07105	141	HS_VehicleTheftsCateg

59	HS_07108	142	Student_CrimesIndex_2020
60	HS_07470	143	Student_PersonalCrimesIndex_2020
61	HS_07501	144	Student_MurdersIndex_2020
62	HS_07502	145	Student_RapesIndex_2020
63	HS_07505	146	Student_RobberiesIndex_2020
64	HS_07514	147	Student_AssaultsIndex_2020
65	Title1_School	148	Student_PropertyCrimesIndex_2020
66	Title1_SchoolWide	149	Student_BurglariesIndex_2020
67	StudentTeacherRatio	150	Student_LarceniesIndex_2020
68	NumFreeLunch	151	Student_VehicleTheftsIndex_2020
69	NumReducedLunch	152	Student_AreaPopulation_2020
70	%_Female	153	Student_Area
71	%_Male	154	Student_TotalCrimesCateg
72	%_EconomDisadv	155	Student_PersonalCrimesCateg
73	%_w/Disabilities	156	Student_MurdersCateg
74	%_ELL	157	Student_RapesCateg
75	%_Homeless	158	Student_RobberiesCateg
76	%_FosterCare	159	Student_AssaultsCateg
77	%_White	160	Student_PropertyCrimesCateg
78	%_Hispanic	161	Student_BurglariesCateg
79	%_Black	162	Student_LarceniesCateg
80	%_Asian	163	Student_VehicleTheftsCateg
81	ALTIERUS CAREER EDUCATION	164	Distance_to_HS
82	BERKELEY COLLEGE	165	Distance_to_PS
83	BLOOMFIELD COLLEGE	166	PS_Graduated

Appendix E. A comprehensive list of the 166 variables in Column Set B.

Appendix F

Variables in Column Set C			
1	Asian	68	ALTIERUS CAREER EDUCATION
2	Black	69	BERKELEY COLLEGE
3	GPA	70	BLOOMFIELD COLLEGE
4	HSGrad_Age	71	CENTRAL CONNECTICUT STATE UNIV
5	Std_07011	72	COLLEGE OF NEW JERSEY (THE)
6	Std_07026	73	COMMUNITY COLLEGE OF RI - WARWICK
7	Std_07103	74	DAYTONA STATE COLLEGE
8	Std_07104	75	DEVRY UNIVERSITY
9	Std_07105	76	DREW UNIVERSITY
10	Std_07106	77	ENTERPRISE STATE COMMUNITY COLLEGE
11	Std_07107	78	ESSEX COUNTY COLLEGE
12	Std_07108	79	FAIRLEIGH DICKINSON UNIV-TEANECK
13	Std_07112	80	FORTIS INSTITUTE
14	Std_07114	81	FRANCIS MARION UNIVERSITY
15	Std_07501	82	KEAN UNIVERSITY
16	Std_07502	83	LINCOLN UNIVERSITY
17	Std_07503	84	MARYMOUNT MANHATTAN COLLEGE
18	Std_07504	85	MIDDLESEX COUNTY COLLEGE
19	Std_07505	86	MONTCLAIR STATE UNIVERSITY
20	Std_07506	87	NEW JERSEY INST OF TECHNOLOGY
21	Std_07513	88	NORTH CAROLINA CENTRAL UNIVERSITY
22	Std_07514	89	OAKWOOD UNIVERSITY
23	Std_07522	90	PASSAIC COUNTY COMMUNITY COLLEGE
24	Std_07524	91	RAMAPO COLLEGE OF NEW JERSEY
25	Absences	92	RIDER UNIVERSITY
26	SAT_Reading	93	ROWAN UNIVERSITY
27	SAT_Math	94	RUTGERS, THE STATE UNIVERSITY OF NJ

28	SAT_Writing	95	SETON HALL UNIVERSITY
29	Honors	96	STANFORD UNIVERSITY
30	Barringer High School	97	STEVENS INSTITUTE OF TECHNOLOGY
31	East Side High School	98	STOCKTON UNIVERSITY
32	Eastside High School	99	SYRACUSE UNIVERSITY
33	John F. Kennedy High School	100	TEMPLE UNIVERSITY
34	Malcolm X Shabazz High School	101	UNION COUNTY COLLEGE
35	Passaic County Technical Institute	102	UNIVERSITY OF NEW HAVEN
36	Rosa L. Parks School of Fine and Performing Arts	103	UNIVERSITY OF WISCONSIN-MADISON
37	School of Earth and Space Science	104	VAUGHN COLLEGE OF AERONAUTICS AND TECHNO
38	School of Government and Public Administration	105	VIRGINIA STATE UNIVERSITY
39	West Side High School	106	WILLIAM PATERSON UNIVERSITY
40	HS_07103	107	HS_CrimesIndex_2020
41	HS_07104	108	HS_PersonalCrimesIndex_2020
42	HS_07105	109	HS_MurdersIndex_2020
43	HS_07108	110	HS_RapesIndex_2020
44	HS_07470	111	HS_RobberiesIndex_2020
45	HS_07501	112	HS_AssaultsIndex_2020
46	HS_07502	113	HS_PropertyCrimesIndex_2020
47	HS_07505	114	HS_BurglariesIndex_2020
48	HS_07514	115	HS_LarceniesIndex_2020
49	Title1_SchoolWide	116	HS_VehicleTheftsIndex_2020
50	NumStudents	117	HS_AreaPopulation_2020
51	NumTeachers	118	HS_Area
52	StudentTeacherRatio	119	Student_CrimesIndex_2020
53	NumFreeLunch	120	Student_PersonalCrimesIndex_2020
54	NumReducedLunch	121	Student_MurdersIndex_2020
55	%_Female	122	Student_RapesIndex_2020
56	%_Male	123	Student_RobberiesIndex_2020
57	%_EconomDisadv	124	Student_AssaultsIndex_2020
58	%_w/Disabilities	125	Student_PropertyCrimesIndex_2020

59	%_ELL	126	Student_BurglariesIndex_2020
60	%_Homeless	127	Student_LarceniesIndex_2020
61	%_FosterCare	128	Student_VehicleTheftsIndex_2020
62	%_White	129	Student_AreaPopulation_2020
63	%_Hispanic	130	Student_Area
64	%_Black	131	Distance_to_HS
65	%_Asian	132	Distance_to_PS
66	%_Hawaiian	133	PS_Graduated
67	%_AmericanIndian		

Appendix F. A comprehensive list of the 133 variables in Column Set C.

Appendix G

Outcome of Each k -Nearest Neighbors Attempt			
Model	Column Set A	Column Set B	Column Set C
<u>Attempt #1</u> Full & Unscaled Dataset; NearestNeighbors; $k = 3$	10	10	10
<u>Attempt #2</u> Full & Scaled Dataset; NearestNeighbors; $k = 3$	14	12	11
<u>Attempt #3</u> Just Predictors & Scaled Dataset; NearestNeighbors; $k = 3$	10	10	10
<u>Attempt #4</u> Just Predictors & Scaled Dataset; NearestNeighbors; $k = 3$	12	12	10
<u>Attempt #5</u> Just Predictors & Unscaled Dataset; KNeighborsClassifier; $k = 1$	6	6	7
<u>Attempt #6</u> Just Predictors & Scaled Dataset; KNeighborsClassifier; $k = 1$	0	0	0
<u>Attempt #7</u> Just Predictors & Unscaled Dataset; KNeighborsClassifier; $k = 2$	7	7	7
<u>Attempt #8</u> Just Predictors & Unscaled Dataset; KNeighborsClassifier; $k = 3$	10	10	10
<u>Attempt #9</u> Just Predictors & Unscaled Dataset; KNeighborsClassifier; $k = 4$	11	11	11

Appendix G. The number of outliers identified by each k -Nearest Neighbors attempt.

Appendix H

Outcome of Each Pruning Attempt			
Model	Column Set A	Column Set B	Column Set C
<u>Attempt #1</u> Full & Unscaled Dataset; NearestNeighbors; $k = 3$; radius = 1.5; algorithm = "kd_tree"; leaf_size = 20;	10	10	10
<u>Attempt #2</u> Full & Unscaled Dataset; NearestNeighbors; $k = 3$; radius = 10; algorithm = "kd_tree"; leaf_size = 5;	10	10	10

Appendix H. The number of outliers identified by each Pruning attempt.

Appendix I

Outcome of Each <i>k</i> -Means Attempt			
Model	Column Set A	Column Set B	Column Set C
<u>Attempt #1</u> Full & Unscaled Dataset	1	1	1
<u>Attempt #2</u> Full & Scaled Dataset	20	20	20
<u>Attempt #3</u> Just Predictors & Unscaled Dataset	1	1	1
<u>Attempt #4</u> Just Predictors & Scaled Dataset	20	20	20
<u>Attempt #5</u> Full & Scaled Dataset; max_iter = 1000	20	20	20
<u>Attempt #6</u> Full & Scaled Dataset; max_iter = 1000; tol = 0.00001	20	20	20
<u>Attempt #7</u> Full & Scaled Dataset; n_init = 20	20	20	20
<u>Attempt #8</u> Full & Scaled Dataset; algorithm = "full"	20	20	20

Appendix I. The number of outliers identified by each *k*-Means attempt. The best-performing attempt is highlighted in green.

Appendix J

Outcome of Each Agglomerative Attempt			
Model	Column Set A	Column Set B	Column Set C
<u>Attempt #1</u> Full & Unscaled Dataset; affinity = "euclidean"; linkage = "complete"	1	1	1
<u>Attempt #2</u> Full & Unscaled Dataset; affinity = "euclidean"; linkage = "single"	1	1	1
<u>Attempt #3</u> Full & Unscaled Dataset; affinity = "euclidean"; linkage = "average"	1	1	1
<u>Attempt #4</u> Full & Scaled Dataset; affinity = "euclidean"; linkage = "average"	1	1	1
<u>Attempt #5</u> Just Predictors & Unscaled Dataset; affinity = "euclidean"; linkage = "average"	1	1	1
<u>Attempt #6</u> Just Predictors & Scaled Dataset; affinity = "euclidean"; linkage = "average"	1	1	1
<u>Attempt #7</u> Full & Unscaled Dataset; affinity = "l1"; linkage = "average"	1	1	1
<u>Attempt #8</u> Full & Unscaled Dataset; affinity = "l1"; linkage = "single"	1	1	1
<u>Attempt #9</u> Full & Unscaled Dataset; affinity = "l1"; linkage = "complete"	1	1	1
<u>Attempt #10</u> Full & Unscaled Dataset; affinity = "l2"; linkage = "average"	1	1	1

<u>Attempt #11</u> Full & Unscaled Dataset; affinity = "manhattan"; linkage = "average"	1	1	1
<u>Attempt #12</u> Full & Unscaled Dataset; affinity = "cosine"; linkage = "average"	33	33	33
<u>Attempt #13</u> Full & Scaled Dataset; affinity = "cosine"; linkage = "average"	45	45	45
<u>Attempt #14</u> Full & Unscaled Dataset; affinity = "cosine"; linkage = "single"	7	7	7
<u>Attempt #15</u> Full & Scaled Dataset; affinity = "cosine"; linkage = "single"	1	1	1
<u>Attempt #16</u> Full & Unscaled Dataset; affinity = "cosine"; linkage = "complete"	36	36	36
<u>Attempt #17</u> Full & Scaled Dataset; affinity = "cosine"; linkage = "complete"	48	45	44
<u>Attempt #18</u> Just Predictors & Scaled Dataset; affinity = "cosine"; linkage = "complete"	45	45	45

Appendix J. The number of outliers identified by each Agglomerative attempt. The best-performing attempt is highlighted in green.

Appendix K

Outcome of Each Logistic Regression Attempt			
Model	Column Set A	Column Set B	Column Set C
<u>Attempt #1</u> Unscaled Dataset	7	6	6
<u>Attempt #2</u> Scaled Dataset	9	9	9

Appendix K. The number of dropouts identified by each Logistic Regression attempt. The best-performing attempt is highlighted in green.

Appendix L

Outcome of Each Stochastic Gradient Descent Attempt			
Model	Column Set A	Column Set B	Column Set C
Attempt #1 Unscaled Dataset	20	20	20
Attempt #2 Scaled Dataset	13	11	12
Attempt #3 Scaled Dataset; penalty = "l1"; max_iter = 1000; tol = 1e-3	13	10	13
Attempt #4 Scaled Dataset; penalty = "elasticnet"; max_iter = 1000; tol = 1e-3	13	11	12
Attempt #5 Scaled Dataset; penalty = "l2"; alpha = 0.01; max_iter = 1000; tol = 1e-3	13	11	15
Attempt #6 Scaled Dataset; penalty = "l2"; alpha = 0.01	11	10	19
Attempt #7 Scaled Dataset; penalty = "l2"; alpha = 0.01; max_iter = 1000; tol = None	11	10	17
Attempt #8 Scaled Dataset; loss = "log"	14	12	12
Attempt #9 Scaled Dataset; loss = "modified_huber"; max_iter = 1000; tol = 1e-3	14	11	11
Attempt #10 Scaled Dataset; loss = "perceptron"	13	11	12

<u>Attempt #11</u> Scaled Dataset; loss = "log"; penalty = "l1"; max_iter = 1000; tol = 1e-3	13	10	11
--	----	----	----

Appendix L. The number of dropouts identified by each Stochastic Gradient Descent attempt. The best-performing attempt is highlighted in green.

Appendix M

Outcome of Each Support Vector Machine Attempt			
Model	Column Set A	Column Set B	Column Set C
Attempt #1 Unscaled Dataset; StandardScaler; LinearSVC; C = 1; loss = "hinge"	15	11	13
Attempt #2 Unscaled Dataset; StandardScaler; LinearSVC; C = 1; loss = "hinge"	10	12	10
Attempt #3 Unscaled Dataset; StandardScaler; LinearSVC; C = 1; loss = "squared_hinge"	15	11	15
Attempt #4 Scaled Dataset; StandardScaler; LinearSVC; C = 1; loss = "squared_hinge"	10	11	11
Attempt #5 Unscaled Dataset; StandardScaler; LinearSVC; C = 10; loss = "hinge"	15	11	15
Attempt #6 Unscaled Dataset; StandardScaler; LinearSVC; C = 50; loss = "hinge"; max_iter = 100000	15	11	15
Attempt #7 Unscaled Dataset; SVC; gamma = "scale"	8	12	8

<u>Attempt #8</u> Scaled Dataset; SVC; gamma = "scale"	7	8	5
<u>Attempt #9</u> Unscaled Dataset; SVC; gamma = "auto"	0	4	0
<u>Attempt #10</u> Scaled Dataset; SVC; gamma = "auto"	0	0	4
<u>Attempt #11</u> Unscaled Dataset; PolynomialFeatures; degree = 3; StandardScaler; LinearSVC; C = 10; loss = "squared_hinge"; max_iter = 100000	14	4	12
<u>Attempt #12</u> Unscaled Dataset; PolynomialFeatures; degree = 3; StandardScaler; LinearSVC; C = 10; loss = "squared_hinge"; max_iter = 100000	5	10	11
<u>Attempt #13</u> Unscaled Dataset; StandardScaler; SVC; kernel = "poly"; degree = 3; coef0 = 1; C = 10	10	6	9
<u>Attempt #14</u> Scaled Dataset; StandardScaler; SVC; kernel = "poly"; degree = 3; coef0 = 1; C = 10	9	8	6
<u>Attempt #15</u> Unscaled Dataset; StandardScaler; SVC; kernel = "rbf"; gamma = 5 C = 0.001	0	8	0

Attempt #16 Scaled Dataset; StandardScaler; SVC; kernel = "rbf"; gamma = 5 C = 0.001	0	0	0
---	---	---	---

Appendix M. The number of dropouts identified by each Support Vector Machine attempt. The best-performing attempt is highlighted in green.

Appendix N

Outcome of Each Decision Tree Attempt			
Model	Column Set A	Column Set B	Column Set C
<u>Attempt #1</u> max_depth = 3	12	9	11
<u>Attempt #2</u> max_depth = 4	12	10	12
<u>Attempt #3</u> max_depth = 2	12	9	9
<u>Attempt #4</u> min_samples_leaf = 3	13	11	10
<u>Attempt #5</u> min_samples_leaf = 4	16	9	9
<u>Attempt #6</u> criterion = "entropy"; max_depth = 3	12	8	9
<u>Attempt #7</u> criterion = "entropy"; min_samples_leaf = 3	13	11	10

Appendix N. The number of dropouts identified by each Decision Tree attempt. The best-performing attempt is highlighted in green.

Appendix O

Outcome of Each Random Forest Attempt			
Model	Column Set A	Column Set B	Column Set C
<u>Attempt #1</u> n_estimators = 100	11	5	9
<u>Attempt #2</u> n_estimators = 200	11	6	11
<u>Attempt #3</u> n_estimators = 500	11	6	13
<u>Attempt #4</u> n_estimators = 300	12	6	8
<u>Attempt #5</u> n_estimators = 200; max_depth = 5	11	6	7
<u>Attempt #6</u> n_estimators = 200; min_samples_leaf = 5	11	7	8
<u>Attempt #7</u> n_estimators = 200; min_samples_leaf = 3	10	8	8
<u>Attempt #8</u> n_estimators = 200; min_samples_leaf = 5; oob_score = True	11	7	8

Appendix O. The number of dropouts identified by each Random Forest attempt. The best-performing attempt is highlighted in green.

Appendix P

Outcome of Each Boosting Attempt			
Model	Column Set A	Column Set B	Column Set C
<u>Attempt #1</u> max_depth = 1; n_estimators = 200; algorithm = "SAMME.R"; learning_rate = 1	15	11	12
<u>Attempt #2</u> max_depth = 1; n_estimators = 200; algorithm = "SAMME.R"; learning_rate = 0.5	15	10	10
<u>Attempt #3</u> max_depth = 1; n_estimators = 200; algorithm = "SAMME.R"; learning_rate = 1.5	15	10	13
<u>Attempt #4</u> max_depth = 1; n_estimators = 200; algorithm = "SAMME.R"; learning_rate = 2	11	13	17
<u>Attempt #5</u> max_depth = 1; n_estimators = 200; algorithm = "SAMME"; learning_rate = 1	15	10	12
<u>Attempt #6</u> max_depth = 3; n_estimators = 200; algorithm = "SAMME.R"; learning_rate = 1	11	7	9
<u>Attempt #7</u> max_depth = 5; n_estimators = 200; algorithm = "SAMME.R"; learning_rate = 1	11	8	9
<u>Attempt #8</u> max_depth = 5; n_estimators = 200; algorithm = "SAMME.R"; learning_rate = 0.54	11	10	12

<u>Attempt #9</u> min_samples_leaf = 3; n_estimators = 200; algorithm = "SAMME.R"; learning_rate = 1	16	7	10
<u>Attempt #10</u> min_samples_leaf = 3; n_estimators = 200; algorithm = "SAMME.R"; learning_rate = 0.5	17	8	10
<u>Attempt #11</u> min_samples_leaf = 3; n_estimators = 200; algorithm = "SAMME.R"; learning_rate = 1.5	16	7	9
<u>Attempt #12</u> min_samples_leaf = 5; n_estimators = 200; algorithm = "SAMME.R"; learning_rate = 1	16	9	9
<u>Attempt #13</u> min_samples_leaf = 2; n_estimators = 200; algorithm = "SAMME.R"; learning_rate = 1	11	7	9
<u>Attempt #14</u> max_depth = 1; n_estimators = 100; algorithm = "SAMME.R"; learning_rate = 0.5	16	9	12
<u>Attempt #15</u> max_depth = 1; n_estimators = 100; algorithm = "SAMME.R"; learning_rate = 1.5	15	8	12
<u>Attempt #16</u> max_depth = 1; n_estimators = 300; algorithm = "SAMME.R"; learning_rate = 1.5	15	9	12

Appendix P. The number of dropouts identified by each Boosting attempt. The best-performing attempt is highlighted in green.

Appendix Q

Outcome of Each Voting Classifier Attempt			
Model	Column Set A	Column Set B	Column Set C
<u>Attempt #1</u> Stochastic Gradient Descent; Support Vector Machine; Boosting; voting = "hard"	17	20	17
<u>Attempt #2</u> Support Vector Machine; Boosting; voting = "hard"	13	4	11
<u>Attempt #3</u> Stochastic Gradient Descent; Boosting; voting = "hard"	15	13	13
<u>Attempt #4</u> Stochastic Gradient Descent; Support Vector Machine; Random Forest; voting = "hard"	15	12	17

Appendix Q. The number of dropouts identified by each Voting Classifier attempt. The best-performing attempt is highlighted in green.