2023

Comparative Analysis of Classification Performance for U.S. College Enrollment Predictive Modeling Using Four Machine Learning Algorithms (Artificial Neural Network, Decision Tree, Support Vector Machine, Logistic Regression)

Anna Kye

LOYOLA UNIVERSITY CHICAGO

COMPARATIVE ANALYSIS OF CLASSIFICATION PERFORMANCE FOR U.S. COLLEGE ENROLLMENT PREDICTIVE MODELING USING FOUR MACHINE LEARNING ALGORITHMS (LOGISTIC REGRESSION, DECISION TREE, SUPPORT VECTOR MACHINE, ARTIFICIAL NEURAL NETWORK)

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE GRADUATE SCHOOL
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

PROGRAM IN RESEARCH METHODOLOGY

BY
ANNA KYE
CHICAGO, IL
MAY 2023

ACKNOWLEDGMENTS

First and foremost, thank you, God, who is always with me every step of the way with countless blessings. Allow me to have a stronger heart whenever I fall and stand up. Amen.

I want to thank my advisor, Dr. Meng-jia Wu, sincerely. I couldn’t complete this long journey without her strong support. Additionally, I would like to express my deepest gratitude to my dissertation committees, Dr. Eilene Edejer and Dr. Ken Fujimoto. Thank you so much. I also want to thank Timothy Heuer, assistant vice president of the Enrollment Systems Research and Reporting Department. Without his advice and support, I couldn’t have considered pursuing a doctoral degree. Finally, I would like to thank my friends and family, who prayed for me with strong encouragement.

To my family.

The heart of man plans his way, but the Lord establishes his steps.

— Proverbs 16:9

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
AUC	Area Under the Curve
BLS	Bureau of Labor Statistics
C	Cost Parameter
CPI	Consumer Price Index
DT	Decision Tree
EFC	Expected Family Contribution
FN	False Negative
FP	False Positive
GB	Gradient Boosting
GDP	Gross Domestic Product
GPA	Grade Point Average
HSAC	High School Academic Climate
HSGPA	High School Grade Point Average
IPEDS	Institutional Postsecondary Education Data System
IRB	Institutional Review Board
KNN	K-Nearest Neighbors
LR	Logistic Regression
ML	Machine Learning

NB	Naïve Bayes
PEL	Parental Education Level
PR	Precision-Recall
RF	Random Forest
ROC	Receiver Operating Characteristic
SES	Socioeconomic Status
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
VIF	Variance Inflation Factor

ABSTRACT

Every year, the national high school graduation rate is declining and impacting the number of students applying to colleges. Moreover, the majority of students are applying to more than one college. This makes a lot of colleges to be highly competitive in student recruitment for enrollment and thus, the necessity for institutions to anticipate uncertainties related to budgets expected from student enrollment has increased. Hence enrollment management has become a pivotal sector in higher education institutions. Data and analytics are now a crucial part of enhancing enrollment management. Through big data analytics-driven solutions, institutions expect to improve enrollment by identifying students who are most likely to enroll in college. Machine learning can unlock significant value for colleges by allocating resources effectively to improve enrollment and budgeting. Therefore, a machine learning method is a vital tool for analyzing a large amount of data, and predictive analytics using this method has become a high demand in higher education. Yet higher education is still in the early stages of utilizing machine learning for enrollment management. In this study, I applied four machine learning algorithms to seven years of data on 108,798 students, each with 50 associated features, admitted to a 4-year, non-profit university in Midwest urban area to predict students' college enrollment decisions. By treating the question of whether students offered admission will accept it as a binary classification problem, I implemented four machine learning algorithm classifiers and then evaluate the performance of these algorithms using the metrics of accuracy, sensitivity, specificity, precision, F-score, and area under the ROC and PR curves. The results from this

study will indicate the best-performed prediction modeling of students' college enrollment decisions. This research will expand the case and knowledge of utilizing machine learning methods in the higher education sector, focused on the U.S. College enrollment management field. Moreover, it will expand the knowledge of how the machine learning prediction model can be pragmatically used to support institutions in setting up student enrollment management strategies.

CHAPTER ONE

INTRODUCTION

Predicting a college's enrollment demographic is a necessity now more than ever. This is because the national high school graduation rate is declining and impacting the number of students who apply to and enroll in colleges (Western Interstate Commission of Higher Education, 2020). Moreover, the majority of students are applying to multiple institutions of higher education (Campbell et al., 2007). This implies that competition is growing among colleges for enrollment and thus the necessity for institutions to anticipate uncertainties related to budgets expected from student enrollment has increased.

Enrollment plays a critical role in the budget and fiscal planning of universities. According to the Integrat-ed Postsecondary Educational Data System (2019), the published cost of attending college for the 2019-20 academic year was $45,543. Even a small shortfall in enrollment, such as ten students, could potentially lead to a financial loss of $455,430 per year for four years, resulting in a total estimated loss of more than $1.83 million. Private colleges are more reliant on revenue generated by students than public colleges, making it crucial for them to accurately predict incoming student enrollment each year (Massa & Parker, 2007). The current study provides valuable insights into identifying students who are more likely to accept admission offers and enroll, which private colleges can use to achieve effective enrollment predictions.

The current study's college prediction model holds substantial importance, has broad applicability, and is easily reproducible. It addresses a crucial issue that almost every institution of higher education faces, namely, "Which admitted students will actually enroll?" Accurate prediction of student enrollment can assist colleges in managing their enrollments effectively, which is a critical aspect of academic administration for many higher education institutions (Hosser & Bean, 1990). The enrollment model's data can be readily obtained by most institutions as part of the admissions process, which makes the current study widely applicable and easily replicable.

The objective of the study is to create a model that can precisely anticipate whether a student will enroll in college by categorizing them into one of two groups: "enrolled" or "not enrolled." To achieve this, the study employed supervised machine learning algorithms, with the student's college enrollment decision problem characterized as a binary classification (Geron, 2017; Hastie et al., 2009). The term "model" refers to the specific algorithm that was chosen after implementing four different machine learning algorithms on the train/validation and test data (Geron, 2017; Hastie et al., 2009).

Implementation of Machine Learning for College Enrollment Prediction Model

The use of accurate predictions allows universities to admit students at the optimal number, preventing over- or under-enrollment (Morgan, 1997). The problem of over-enrollment affects the student experience negatively because it strains institutional resources (Cornell University Division of Planning and Budget, 2006; Zeng et al., 2015). For instance, when a college enrolls too many students, its dormitories and class spaces may not be sufficient to accommodate the students and/or the faculty to teach required general education classes. In contrast, under-enrollment can decrease tuition revenue, result in inefficient use of expanded resources, and reduce the cohort from which universities may seek future donations. Even though it may vary from institution to institution, losses of a few students to enroll can result in million-dollar losses.

It is, therefore, possible for university leaders to meet their admissions and revenue goals by accepting the ideal number of applicants based on how many acceptances will result in enrollments.

The use of data and analytics by higher education institutions is now a crucial part of enhancing their admission and enrollment (Seres et al., 2018). Through big data analytics-driven solutions, institutions expect to improve matriculation by identifying students who are most likely to enroll in college (Antons & Maltz, 2006). In fact, machine learning can unlock significant value for colleges by ensuring that resources are allocated in the most effective ways to improve enrollment and budgeting (Drake & Walz, 2018; Antons & Maltz, 2006). Machine learning combines statistics, mathematics, and computer science into one problem-solving pathway (Dara et al., 2022). This involves discovering patterns, training, and testing data to create computer programs that automatically recognize complex structures and make intelligent decisions (Michell, 1997). Moreover, machine learning algorithms can be implemented in a broader spectrum for compiling prediction models (Delen, 2010; Shabestari et al., 2019; Luan & Zhao, 2006). Hence, the machine learning method is a vital tool for analyzing a large amount of data, and predictive analytics using this method has become a high demand in the higher education sector (Delen, 2010; Shabestari et al., 2019; Luan & Zhao, 2006). Yet higher education is still in the early stages of utilizing machine learning for enrollment management (Dorn et al., 2020).

Machine learning can be utilized to analyze applicants and predict their likelihood (i.e., probability) to accept admission (i.e., enrolling). The machine learning model examines trends among similar applicants from previous years, including multiple metrics such as demographics and academic achievements (e.g., high school GPA and standardized test score). Many previous methods for predicting students’ enrollment decisions relied on formulaic calculations, such as a simple logit regression (Sternberg, 2010). Many formulaic calculations were considered viable since the types of attributes that influence college enrollment decisions and the number of students who attended college were relatively small compared to the current admission pool (Sternberg, 2010). Today, more diverse factors influence students' college enrollment decisions, and the number of students attending college has expanded significantly (Bingham & Solverson, 2016). This has led to the volume of student admission data expanding significantly. Hence, the college enrollment prediction cannot be generated based on the formulaic calculations since those calculations cannot handle a large amount of data and adjust to reflect various factors (Bingham & Solverson, 2016; Shabestari et al., 2019; Luan & Zhao, 2006). Machine learning is therefore utilized to improve the accuracy of the predictions.

In comparison to other approaches employed in prior studies for predicting college enrollment, the Machine Learning (ML) method differs in several ways. Firstly, ML can capitalize on large amounts of data to enhance its predictive capability (Najafabadi et al., 2015). By utilizing ML, it is possible to draw data from various institutional and public sources and combine it to train models that identify patterns and utilize them for making predictions (Najafabadi et al., 2015). With the help of machine learning, an institution can predict student enrollment by integrating historical data with external data, such as the unemployment rate (Ekowo & Palmer, 2016). Because ML models can incorporate data from diverse sources and process vast amounts of information, they can achieve greater accuracy than previous methods. ML is different from conventional, formulaic computation approaches as it involves ongoing cycles of discovery and implementation, utilizing data to explain phenomena and make predictions (Sternberg, 2010; Bingham & Solverson, 2016).

Multiple types of ML algorithms can be utilized to tackle problems such as predicting college enrollment. In the current study, four ML algorithms were evaluated: support vector machine, decision tree, artificial neural networks, and logistic regression. ML algorithms are notoriously data-intensive (Mikolajczyk & Grochowski, 2018). Therefore, in order to make an accurate prediction using ML, it is beneficial to aggregate a pool of data that encompasses not only institutional but also publicly available data that could influence any of the factors related to college enrollment decisions. Relevant public data may include economic, education, and population statistics sourced from the US Bureau of Labor Statistics (BLS). The BLS has been accumulating long-term data on inflation and unemployment rates going back decades, totaling about 100 years of data. All of this data can be employed to train machine learning algorithm models and enhance their predictions about the future. However, before an institution adopts machine learning, it is important to recognize that machine learning-based prediction tools should only be used to offer more precise data to higher education admission stakeholders, who will remain the ultimate decision-makers concerning acceptance (Rodríguez-Muñiz et al., 2019). By revamping existing methods and improving data accuracy, machine learning can furnish better insights to decision-makers and aid them in formulating optimal enrollment management strategies for their institutions.

Proposed Analysis

The current study focuses on seven years (2013 to 2019) of de-identified admission data from a particular college and deploys four different machine learning algorithms (i.e., support vector machine, artificial neural network, decision tree, logistic regression) for creating students' enrollment prediction models. The attributes reflected in the prediction models are identified based on the college choice (i.e., enrollment) conceptual model proposed by Perna (2006) and other relevant literature. Identified attributes that significantly impact students' college enrollment decisions are related to students' habitus (e.g., gender, race, socioeconomic status, academic performance), high school and community context (e.g., parental education level, secondary school academic climate), higher education context (e.g., college proximity, institutional financial aid), and social, economic, and policy context (e.g., national unemployment rate, inflation rate).

In addition to identifying influential factors on students' college choice decisions based on Perna's (2006) framework, I chose four different machine learning algorithms based on pertinent studies. These past studies conducted comparative analyses using multiple machine learning algorithms for predicting U.S. college enrollment (Walczk & Sincich, 1999; DesJardins & Gonzales, 2002; Gerasimovic & Bugaric, 2018; Antons & Maltz, 2006; Chang, 2006; Vialardi et al., 2011; Ragab et al., 2014; Cirelli et al., 2018; Slim et al., 2019; Lux et al., 2020). The conclusion of the best machine learning algorithm selected for college enrollment prediction modeling differed across all studies. However, no research has since conducted a comparative analysis reflecting the machine learning algorithms that were chosen to be the best by past studies. Hence, the four different machine learning algorithms (i.e., logistic regression, decision tree, support vector machine, and artificial neural network) were chosen because they were identified to be the best to predict U.S. college enrollment.

To compare prediction models based on four different machine learning algorithmic methodologies, four main components of measuring the model fit (i.e., accuracy, sensitivity, specificity, precision) were calculated (Brieman et al., 1984). In addition, F_1 score (Chicco &

Jurman, 2020) was calculated and compared. To assess the diagnostic performance of prediction, I plot the Receiver Operating Characteristic (ROC) curve and calculate the Area Under the ROC Curve (AUC) for each model (Fawcett, 2006). I also preprocessed the data using K-fold cross-validation to verify that the model results are not biased. All the above-mentioned performance metrics were used to compare the models to find the one that is best suited for enrollment prediction based on the college admission data.

Research Questions

The purpose of college enrollment predictive modeling is to apply the logic of past admitted students' actions who enrolled/not enrolled to a future group of students by consistently comparing the same metrics in a similar college choice framework. This study aims to generate four different prediction models, investigate which college choice factors are most influential across models and determine the best prediction model in terms of statistical validity. Thus, the following questions are explored:

1. Which machine learning prediction model (Logistic Regression vs. Artificial Neural Network vs. Decision Tree vs. Support Vector Machine) demonstrates the best prediction performance for testing Fall 2019 enrollment after training the historical data (from Fall 2013 to Fall 2018)?
2. With a machine learning prediction model chosen in Q1, what are the most significant factors contributing to the prediction of the enrolled students for Fall 2019?
3. With a machine learning prediction model chosen in Q1, what applications can be followed up and utilized to support a higher education institution's enrollment management strategy as a plan of action?

Significance of Study

Identifying the best college enrollment prediction model among four machine learning algorithms is not intended to provide a definitive answer to whether a student will enroll. There are too many variables to consider, and the answer will be known eventually as time goes by. However, the best prediction model will provide a solid method to use for deciding which students are highly likely to enroll based on the attributes reflected in the model. Since such a model offers the enrollment probabilities for admitted students, those probabilities will then inform the admission department on how best to invest its limited recruitment resources.

In this era of maximizing resources, the best prediction model will allow colleges to target recruitment efforts to admitted students who are undecided, follow up with those who are likely to enroll and/or reconfirm their commitment and decision and create a better understanding of the applicant pool, including the characteristics of students' demand to be a good match for the college.

In summary, I discussed a problem that I consider crucial in chapter one and how it has been built up. As a senior research & strategy analyst at the enrollment division in a higher education institution, developing and compiling reliable college enrollment prediction model(s) is very important and eventually led me to write a dissertation.

CHAPTER TWO

LITERATURE REVIEW

The enrollment management process is directly involved in new student recruitment, financial aid planning and budgeting. To maximize student enrollment, enrollment management administrators should frequently communicate with the admissions office and financial aid department to set up better plans on student recruitment along with allocating institutional scholarships and grants. Considering the large expenditure of institutional marketing and financial aid awards, this research study explores different attributes that influence college choice (i.e., enrollment) of admitted students and implements machine-learning algorithm models for predicting admitted students' enrollment decisions.

The objectives of the current study are 1) to provide a better understanding of both student and institution-related attribute(s) that highly influence students' enrollment decisions and 2) to identify the best predictive model(s) using machine learning algorithms that institutions can discern admitted students who are highly likely to enroll relatively to others. This allows enrollment management to support financial aid, admission, and marketing departments to assign budgets effectively and thus not only maximize enrollment but also save a big portion of marketing efforts efficiently for recruiting students. In this chapter, I discuss the factors and enrollment prediction models using machine learning algorithms, including a wide range of features related to student and institutional characteristics.

Background of College Choice Model

College choice theories generally center on three major questions: 1) who goes to college, 2) where do they enroll, and 3) why do they select that specific college? These college choice theories draw upon various disciplines to establish their studies on theories from a broad range of perspectives. While numerous college choice models have been created based on the three questions, they have more recently been classified into one of three primary subgroups: economic, psychological, or sociological (Bateman & Spruill, 1996; Hossler & Palmer, 2008; McDonough, 1997; Stage & Hossler, 1989). Despite the distinct disciplinary perspectives and specific areas of focus within college choice theories, many of these paradigms complement each other in the process of selecting a college.

College choice models that adopt an economic perspective regard selecting a college as a type of decision-making behavior that resembles an investment (Jackson 1978). From an economic standpoint, a rational decision-making process is primarily based on tuition fees, available resources, and financial assistance (Archibald & Feldman, 2010; Bateman & Spruill, 1996; Hossler & Palmer, 2008; Paulsen & St. Johnson, 2002; St. John et al., 2010). These models depend on the cost and availability of resources to demonstrate what factors influence a student’s ultimate decision. An economic approach known as the input model considers money as the ultimate influence and disregards any external personal or social factors that may affect the decision (Bateman & Spruill, 1996; Hossler & Palmer, 2008; Paulsen & St. Johnson, 2002). Another economic approach is the output model, which bases college choice decisions on the potential financial benefits after graduation and considers institutional prestige and disciplinary options in the decision-making process (Archibald & Feldman, 2010; Bateman & Spruill, 1996; Hossler & Palmer, 2008; Paulsen & St. Johnson, 2002; St. John et al., 2010). Within the context of choosing a college, economic perspectives encompass theories that emphasize financial factors, as well as sociological theories that take into account a combination of monetary benefits and social standing. Economic models serve as a crucial factor in the decision-making process for college choice across various models.

Several other models for college choice utilize psychology as a foundation to comprehend the various elements that influence students' decisions concerning college choice. The emphasis of psychologists lies in the psychological climate or environment of an educational institution, its effects on students, and the compatibility of students with the institution (Astin, 1965). These models often examine the influences of others (friends, family, and counselors) whom students build a relationship in their community and the academic climate that is offered by the pre-college institutions where they attend (Hossler & Gallagher, 1987; Hossler & Palmer, 2008) on student decision-making. Chapman's (1981) approach to the college choice model was distinctive, as it anazlyed the background ahd personal traits of students with respect to the communities they are associated with. He determined the influential weight of these variables on the student's college decision process (Chapman, 1981).

A third perspective to investigate college choice is grounded in sociology. Sociologists perceieve the development of aspirations toward higher education as an aspect of a broader process of achieving social status (Deil-Amen & Turley, 2007). For instance, since the 1960's, the social attainment model has been a significant paradigm for studing educational and career aspirations (Kao & Tienda, 1998). It demonstrated how the convergence of family background and resources influences a child's upbringing and, eventually, their educational aspirations (Kao & Tienda, 1998). McDonough (1997) was motivated by Bourdieu's (1986) assessment of external factors that influence an individual's decision-making process and explored how a student's social class determines suitable option for college choice. This study also illustrated that each high school has a distinct set of established values and social standards that could determine which college choice are deemed acceptable by the student's peer group for application and enrollment (McDonough, 1997).

Gaining insights into the diverse disciplinary viewpoints, such as economics, psychology, and sociology, concerning college choice theories establishes a fundamental framework for comprehending the numerous factors that impact decision-making for college choice.

College Choice Models

College choice models illustrate the various routs, phases, and significant factors that influence a student's decision-making process for selecting and enrolling in a college. Enhancing the comprehension of college choice facilitates families, high school counselors, and higher education stakeholders to gain a better understanding of the primary factors that hold the most influence in the college choice process. They also can utilize those factors to better strategize institutional financial budgets and student recruitment for next academic cycle. Moreover, the college choice model can be expanded to make predictions on student enrollment from various perspectives, which support the idea of designing the future student body.

Kinzie et al.'s (2004) study on the history of higher education revealed the earliest models of college choice that originiated in the 1950s. According to Kinzie et al. (2004), the 1940s established the groundwork for increased access to higher education through the introduction of the Servicemen's Readjustment Act ("GI Bill") of 1944 and President Harry S. Truman's initiatives to expand community college systems. Soon after these significant developments, the U.S.

Supreme Court made a decision in Brown v. Board of Education of Topeka (1954). In the aftermath of the Brown v. Board of Education ruling, endeavors were initiated to integrate public schools nationwide and increase the accessibility of college education for minorities. These efforts augmented the college-going populace and compelled colleges to devise a more advance procedure where prospective students had to make enrollment decisions about which college to attend (Kinzie et al., 2004).

Holland (1959) was the pioneer to publish research on college choice. Using data from 814 elite high school students, he conducted an empirical model of college choice. In this study, Holland (1959) found that students' decisions were influenced by the interplay of various characteristics, including student and parental interests, attitude, educational background, gender, and socioeconomic status. However, he identified the complexity of the college selection process and demonstrated that students of various backgrounds select different kinds of institutions (Holland, 1959). He also stated, "like many personal decisions, the choice patterns found here are probably not really amenable to change because they are grounded in cultural and personal development" (Holland, 1959, p. 26). Since then, there have been many studies exploring the college choice process and influential factors.

Later scholars, including Kotler, leveraged the research of Holland (1959) in order to demonstrate the college choice process. The first model of college choice was developed by Kotler (1976). He applied consumer decision-making process theory to see how students select colleges to enroll, which treats them as consumers (Kotler, 1976). This theory encompasses the concepts of the students' behaviors in selecting the college before applying and three major questions which were mentioned previously: 1) who goes to college, 2) where do they enroll, and 3) why do they select that specific college? Kotler (1976) distinguished seven stages of the college decision process: 1) making a decision to attend college, 2) seeking and receiving information about colleges, 3) submitting inquiries to specific colleges, 4) submitting applications, 5) obtaining acceptance letters, 6) making a college choice, and 7) registering for classes. In addition to proposing seven stages of the college choice process, Kotler (1976) stressed the significance of the institution in the ulticmate college choice decisions. While Kotler’s model was widely accepted as accurate during his time, more recent research helped refine his theory. Hanson and Litten's (1982) model re-evaluated Kotler's (1976) model, creating a five-stage model that comprised 1) college aspirations, 2) search, 3) information gathering, 4) application, and 5) enrollment. Following the Hanson and Litten (1982) five-stage model. Researchers continued to simplify the college decision choice model further, distilling the process into a basic three-phase model.

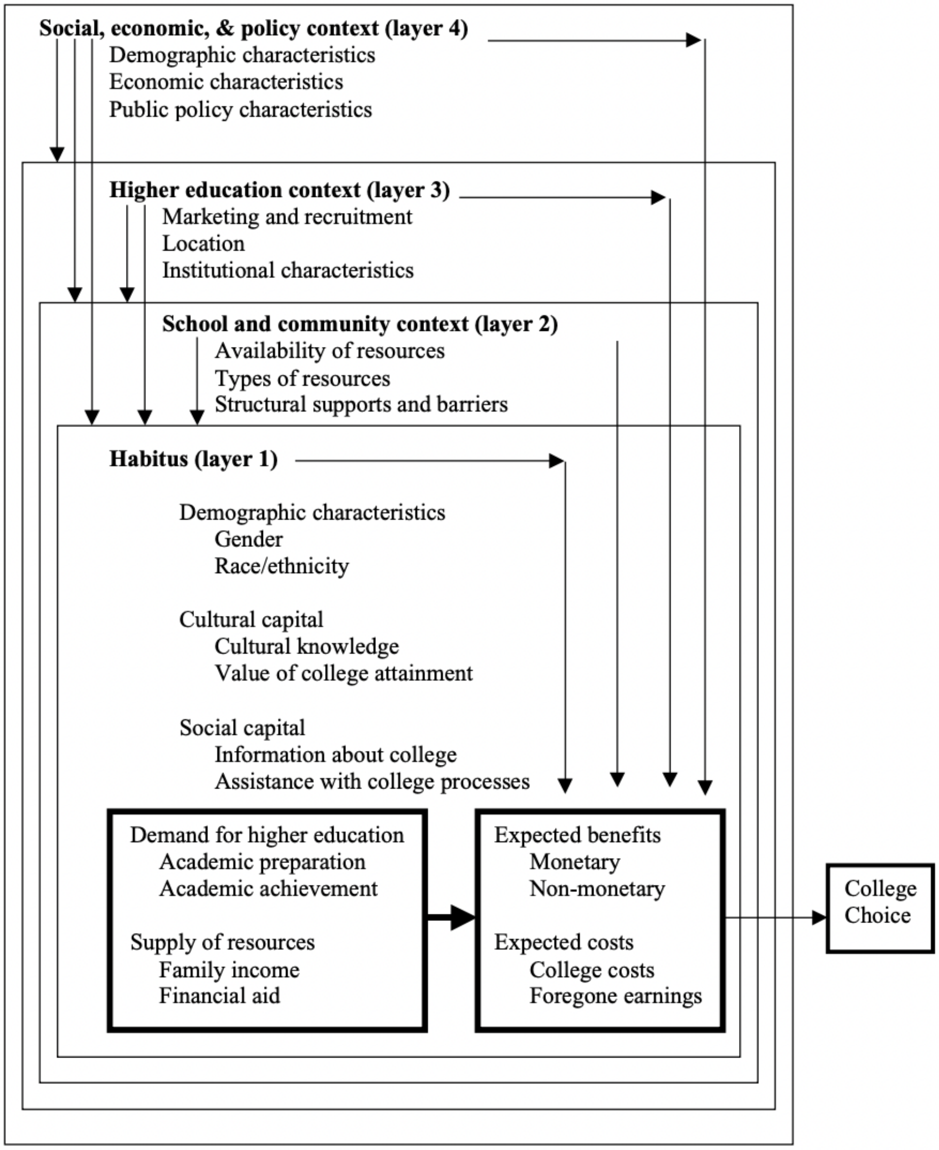
Jackson (1986) exapanded on Hanson and Litten’s (1982) model by utilizigin data gathered in a longtitudial study, ultimately constructing three categories of the college choice process. The first category is called the preference phase. It indicated that students’ educational aspirations are highly correlated with their academic achievement which are influenced by the students’ family background and social context. The second category, called the exclusion phase, involves students excluding some institutions from the prospective list based on the available resources. Potential resources that result in exclusion are college tuition, fees, locations, and/or academic quality. After going through the exclusion phase, the third category is the evaluation phase. In this phase, students limit their college choice and finalize a list of college to choose.

Apart from to Jackson’s (1986) model, Hossler and Gallagher (1987) formulated another three-stage model. The initial stage referred to as pre-disposition, which recognizes the period when a student deteremines whether to attend college (Hossler & Gallagher, 1987). After students make the decision to attend college, they progress to the second stage of decision-making which is the search process (Hossler & Gallagher, 1987). During the search process, students start gathering information about various colleges through formal and informal means, and they start making emotionally driven decisions regarding specific institutions of higher education (Hossler & Gallagher, 1987).

The last stage is choice stage, during which students make their final decision. This college choice model also take into account various additional factors such as location, availalibity of financial aid, academic quality, campus visits. Furthermore, the model highlights the ways in which these factors play a role in shaping student’s ultimate decision regarding college choice (Hossler & Gallagher, 1987). Further research is needed to determine the extent of influence of each factors in college choice research, despite the continued use of college choice models from the 1980s in current counseling practices (Hossler & Palmer, 2008). In addition to a need for more research about each factor within the three-stage model, there also has been a push for more research to expand outside of the basic three-stage model (Jackson, 1986; Hossler & Gallagher, 1987).

Research on college choice continued in the last two decades with an interest in the factors impacting students' college choice decisions. These studies developed five- and three-stage models but also found additional new factors that appear to significantly influence on students’ enrollment decisions. Perna (2006) proposed a college choice conceptual design using multilevel modeling which contains four layers: 1) student's habitus, 2) school and community context, 3) higher education context, and 4) social economic, and policy context. Perna (2006) states that a multi-level model that is used here addresses the hierarchical relationship of factors that are grouped into four different layers. Moreover, Perna (2006) illustrated there are additional factors that are not covered in the three-stage model which is related to social, economic and policy contexts. Figure 1 shows her conceptual model on how students' college choice decision-making is affected by factors in four different layers.

Figure 1. Perna's (2006) Proposed Conceptual Model of Student College Choice


Social, economic, & policy context (layer 4)
Demographic characteristics
Economic characteristics
Public policy characteristics
Higher education context (layer 3)
Marketing and recruitment
Location
Institutional characteristics
School and community context (layer 2)
Availability of resources
Types of resources
Structural supports and barriers
Habitus (layer 1)
Demographic characteristics
Gender
Race/ethnicity
Cultural capital
Cultural knowledge
Value of college attainment
Social capital
Information about college
Assistance with college processes
Demand for higher education
Academic preparation
Academic achievement
Supply of resources
Family income
Financial aid
Expected benefits
Monetary
Non-monetary
Expected costs
College costs
Foregone earnings
College Choice

Note: Proposed conceptual model of student college choice. Adapted from Studying college access and choice: A proposed conceptual model by Perna, L. (2006) in Higher Education: Handbook of Theory and Research. Vol. XXI, p. 117.

The first level of Perna’s (2006) model is the student’s habitus. The habitus reveals “an individual’s demographic characteristics, particularly gender, race, and socioeconomic status, as well as cultural and social capital” (Perna, 2006, p. 117). Perna regarded this layer as crucial in the decision-making process as it focuses on the individual student and their unique characteristics that are ingrained and persistant over prolonged period.

The subsequent three layers, excluding the first, are contextual and comprise external factors that affect the students. However, these are potentially influential to students’ college enrollment decisions. These three contextual layers include the school and community context; the higher education context; and the social, economic, and policy context. The specific descriptions for the rest three layers are stated in the following paragraphs.

The second layer is the secondary school and community context of college choice. This layer centers on how schools and communities impact students’ college choice decisiosn. Perna (2006) suggested that schools can influence students’ college choice decisions in various ways, such as providing teacher encouragement, offering college preparation courses, and promoting high-quality extracurricular activities. In addition to secondary school, Perna (2006) claims that communities significantly influence students’ college choice decisions.

Perna’s (2006) proposed that higher education institutions influence students’ enrollment decisions in three ways within the third layer of her conceptual model. Firstly, colleges provide information to prospective students and families through activities such as college visits. Secondly, instituions’ attributes and characteristics, such as proximity and the availability of institutional financial aid, play a role. Finally, the availability of enrollment slots at the institution can also impact students’ college choice.

Finally, according to Perna (2006), the outermost layer of her conceptual modell, which is influenced by “social forces (e.g., demographic changes), economic conditions (e.g., unemployment rate), and public policies (e.g., the establishment of a new need-based grant program)” (Perna, 2006, p. 119), plays a cruicial role in the decision-making process, particularly in the social, economic, and policy context of the decision.

Perna’s (2006) model incorporates multiple layers, which assume that the college choice decision-making is influenced by multiple factors. These layers propose that the decision to enroll in college is based on comparison of the benefits and costs of enrolling and that this assessment is shaped by an individual’s habitus, as well as the family, school, and community context, higher educationa context, and social, economic, and policy context.

This study is based on Perna’s (2006) model because it provides a comprehensive framework that takes into account multiple layers of influence that interact to impact stduents’ college choice, thereby offering a complex understanding of the decision making process. Additionally, it contains layers that include not only factors involved in three-stage models but also involves new factors that encompass broader spectrums of social forces, economics, and policies. Since the current study specifically focuses on the admitted students’ enrollment decision process, it reflects the partial phase of three-stage models. However, three-stage models present related factors as discrete by each phase. Hence, Perna’s multi-layer model, which all factors are nested across the enrollment decision process, is appropriate to implement as a conceptual framework. This study views the familiar variables of college enrollment decision-making and looks specifically at one of the 4-year private Jesuit universities in the mid-west urban area. The study concludes that professionals at the college admissions level could benefit from the information provided by college choice model(s).

Influential Factors of College Choice/Enrollment

Within the college choice models and theories, ten significant factors of influential factors consistently arise in the current research, which are student's 1) gender, 2) race, 3) socioeconomic status (e.g., Expected Family Contribution), 4) academic performance (e.g., High School GPA, ACT/SAT scores), 5) parent education level, 6) secondary school academic climate, 7) college proximity, 8) institutional financial aid, 9) national unemployment rate, and 10) national inflation rate. Since this study proposes to use Perna's college conceptual model, Table 1 shows how these ten factors are assigned to each of four different layers in terms of (a) students' habitus, (b) high school and community context, (c) higher education context, and (d) social economic, and policy context.

Table 1. Influential College Choice Factors Assignment Based on Perna's College Choice Model

Students' habitus	High school and community context	Higher education context	Social, economic, and policy context
* Gender * Socioeconomic Status (i.e., Expected Family Contribution) * Race * Academic Performance (ACT, SAT, High School GPA)	* Parent Education Level * High school academic climate (i.e., College Board's Secondary school segment rating)	** College proximity* ** Institutional financial aid*	** National unemployment rate* ** Inflation Rate (i.e., Customer Price Index * 100)*

Note. Adapted from "Studying college access and choice: A proposed conceptual model" in Higher Education: Handbook of Theory and Research, Vol 21 (p. 117), by L. W. Perna, 2006, Springer, Netherland: Dordrecht

Students' Habitus Characteristic Context Factors

College choice models can be used to explain the decision-making process of students, as well as their demographic characteristics and academic achievements, all of which are important factors in the college choice process of high school seniors. During the application process, applicants typically provide demographic information, such as gender, race, socioeconomic status (e.g., Expected Family Contribution), and academic achievements (e.g., High School GPA, ACT/SAT scores). Enrollment managers compile the data from the applications to discover demographic information and academic performance of applicant pools. Understanding the importance of demographic factors is imperative for higher education instituiosn. This information strongly connects with a high school seniors' college decision-making process (Kim, 2004), and institutions can find ways to identify and address anything in their profiles that could be appealing to a specific demographic group (Horvat et al., 2003).

Gender differences in college enrollment decisions have been tracked and documented for decades (NCES, 2020; Carbonaro et al., 2011; Flashman, 2013; Kleinfeld, 2009; Reynolds & Burge, 2008; Turley et al., 2007; Fortin, Oreopoulos, & Phipps, 2015). In 1960, over 60% of college enrollees were men (NCES, 2020). However, the rate of women's college enrollment increased substantially over the next two decades (NCES, 2020). For the gender gap in college enrollment, some scholars pointed to several explanatory factors. According to Carbonaro et al (2011), females in high schools are more likely to have higher academic achievement and aspirations than males and this may explain why females are more likely to apply and enroll in college by their senior year of high school. Flashman (2013) and Kleinfeld (2009) corroborated Carbonaro et al.'s (2011) assertion that females' superior academic performance may have contributed to gender disparities in postsecondary participation, particularly among low-income students. However, male students have been found to have lower academic aspirations, are less likely to enroll in college prep courses (Reynolds & Burge, 2008), and apply to college during their senior year of high school (Carbonaro et al., 2011; Turley et al., 2007). This may be because males are more inclined to pursue postsecondary plans that do not involve college, such as joining the military or attending vocational schools (Fortin et al., 2015; Reynolds & Burge, 2008). In other words, this suggests that males are less likely to view postsecondary education as essential for their future employment opportunities (Carbonaro et al., 2011; Kleinfeld, 2009).

Since higher education is facing an increase of gender inequality, many policy makers, researchers, and educators are seeking ways to minimize and balance the gender gap of the student body. For instance, some colleges' administrators and admission officers expressed concerns about providing advantages to female applicants to enroll and advocated implementing affirmative action for male applicants (Kao & Thompson, 2003; Greene & DeBecker, 2004).

As mentioned earlier, another important influence on college decision-making in the student habitus is socioeconomic status (SES), which is generally measured by parental education and household income (Attewell & Domina, 2008; Handwerk et al., 2008). Adelman (2006) found that students' SES was significantly related to their transition to college; higher-SES students attended college more often than those with lower SES. It is common for students to make admissions decisions based on their social class and the community in which they live. When students begin discussing college, factors based on SES are prevalent (Kablenberg, 2004; McDonough, 1997). For instance, those from middle-class families often receive information about college from relatives who have attended college (Bloom, 2007). By contrast, students from low-income families tend to protect themselves from the reality of rejection and do not discuss college plans as often with their families (Bloom, 2007). McDonough (1997) asserts that students' attitudes toward college enrollment decisions are influenced by the expectations of family members and society. Gladiuex (2004) studied the relationship between students' SES levels and college enrollment decisions, including their high school academic achievement using standardized test scores. Gladieux (2004) explained that students from lower-SES backgrounds attended college in much lower numbers than students from higher-SES backgrounds, regardless of their academic achievement in high school. Furthermore, he observed that students in the highest-SES quartile range but who scored in the lowest test score quartile range were more likely to attend college than those from the lowest-SES quartile range but who score in the highest-test score quartile range. This implies that the least smart, rich kids have as good a chance of going to college than the smartest, poor kids (Gladiuex, 2004). In the current study, expected family contribution (EFC) is used o measure the students' SES since EFC directly reflects the students' financial strength.

Besides differences in gender and SES, racial groups also differ in the extent of their college enrollment decisions (Adelman, 2006; McDough, 1997). Many racial differences in educational achievement can be partially accounted for by including family background and SES measures (Beattie, 2002). There are significant racial gaps in grades and test scores, thus some racial minority high school graduates may be hindered in their attempts to attend college (Roscigno, 2000). Adelman (2006) found that knowledge of and attitudes toward college often vary with race. He further noted that despite the increased participation of racial minority students in postsecondary education over the previous quarter century, the gap in college enrollment between Whites and Asians vs. Hispanics and African Americans remained wide. According to the National Center for Education Statistics (NCES), there were 16,610,200 students enrolled in higher education institutions in fall 2018; Caucasians made up 54% of this group, while African Americans comprised 13.3% of the undergraduate population (NCES 2020). While more students than ever are going to college, the disparity between races persists.

As mentioned earlier, academic achievement measures like standardized tests (i.e., ACT/SAT) and high school GPA are significant predictors of college enrollment (Cho, 2007; Klasik, 2012). In general, submitting standardized test scores and high school GPA (HSGPA) are required in the admissions processes of all four-year colleges and universities. Cho (2007) examined high school GPA as a factor in college entrance patterns and found that it was a highly important determinant of attending college, in addition to standardized test scores. Allensworth and Clark (2020) also stated that there is a strong correlation between HSGPA and college enrollment. They stated that students with higher HSGPA are more likely to enroll in college since they have strong educational aspirations (Allensworth & Clark, 2020). Similar to HSGPA, Klasik (2012) showed that taking the ACT or SAT is a major step to college enrollment among high school students. He found that 95% of students who enrolled in a four-year college or university took the ACT and/or the SAT at some point between 10th and 12th grade (Klasik, 2012). In fact, taking standardized tests is one of the most predictive steps of later college application and enrollment.

High School and Community Context Factors

Many factors beyond demographic characteristics and academic achievement affect students' college choice processes. When making college choice decisions, students are inundated with various messages coming from persuasive sources, including families and communities (Attewell & Domina, 2008; Handwerk et al., 2008). The main social influences are parental education level (Cameron & Heckman, 2001; Belzil & Hansen, 2003) and a student's high school academic climate (Engberg & Wolniak, 2010; Nuñez & Kim, 2012). Hence, these two main sources are used in the present study.

Dornbusch, et al. (1987) claimed that parental education level affects children's academic path, especially for their college journey. The study by Cameron and Heckman (1998) also showed that parents' education level was by far the most important family background variable for students' college education attainment. These two factors account for as much as 83% of the explained variations in the student's college attainment outcome (Cameron & Heckman, 1998). Parents with a high school diploma and students who are the first in their immediate family to enroll in college have lower participation rates in academic programs to prepare for college enrollment and lower rates of applying to and enrolling in college (Horn & Bobbitt, 2000). In addition, a number of studies stated that students whose parents have gone to college are more likely to attend college themselves (Goyette, 2008; Bifulco et al., 2011; Choi et al., 2015). As mentioned earlier, students' SES is generally measured by parental education level (Attewell & Domina, 2008; Handwerk et al., 2008). Students with higher SES attend college more than often than those with lower SES. Stange (2012) indicated that their predicted lifetime income increases with a higher parental education level, affecting their children's academic aptitude, eventually leading to their college choice and enrollment. The study by Belzil and Hansen (2003) also showed a positive correlation between individual schooling attainment (e.g., college enrollment) and parents' education level.

Secondary school is also a significant community where students are involved in. Hence, the secondary school academic climate has become increasingly popular in discussions of college choice influential factors (Astin et al., 2011; Cohen et al., 2009; Zulling et al., 2010). Specifically, creating "college-going cultures" in secondary schools has been the focus of efforts to expand postsecondary access (Knight & Marciano, 2013; Knight et al., 2019). Generally, college-going culture is highly related to students' academic preparation (Conley, 2012; Duncheon, 2015; Hooker & Brand, 2010). Rigorous academic preparation has been cited as a leading predictor of college success (Adelman, 2006; Perna, 2005; Porter & Polikoff, 2012). Students' academic preparation has typically been addressed through the use of test scores, course levels, and other standardized measures of student achievement in the college and career readiness process (Wearne, 2018). Therefore, high schools strive to enhance students' academic readiness, especially for college attendance, by increasing advanced coursework offerings, such as AP courses and dual credit classes, as well as holding students to high academic standards (Jarsky, et al., 2009). Due to the fact that this metric is significantly influential on students' college enrollment decisions, the present study includes the part of College Board database which rate all U.S. high schools based on 40 academic and demographic factors (see Appendix B for more detailed information).

Higher Education Institution Context Factors

In addition to the influence of secondary schools' academic climate on students' college choice, higher education institutions have a direct impact over certain specific factors in students' college choice decisions. They include the geographic proximity between students' hometown and college, and institutional financial aid such as scholarships and grants. There are other possible factors such as the level of students engagement with college but that is not covered since the way how engagement data are defined and classified are highly subjective.

According to Chute (2006), one of the most critical institutional factors influencing a student's college choice is the proximity of the institution to their hometown. The research found that 56% of students attend a higher education institution located within 100 miles of their hometown. Turley (2009) discovered that location was one of the primary reasons for selecting an institution, with proximity to home being the most significant influence on that decision. In other words, students were more likely to apply to and attend institutions that were comparatively close to their homes. Disadvantaged students, in particular, viewed nearby institutions as their only practical option for higher education, allowing them to save money on room and board by living at home. On the other hand, Hoxby (2009) discovered that improvement in transportation had raised the probability of students feeling at ease with the idea of attending a college or university located far away from their residence. Despite that result, advances in transportation, proximity remains a crucial factor for many high school seniors in their college decision-making process.

Along with proximity, students' college choices are heavily influenced by institutional financial aid. Many forms of institutional financial aid are available, including scholarships, grants, and merit aid. Since college financial aid helps students cover tuition, fees, boarding, books, supplies, and so on, thus aid plays a vital role in students' enrollment decisions. According to Farrell and Kienzl (2009), states that provided generous institutional financial aid such as merit awards saw the greatest increases in college enrollment. In states that offered the top quartile of merit aid amounts (i.e., nearly full-tuition scholarships), college participation increased by 5.5%, and the enrollment rate among freshmen staying in that state for college increased by 6.6% between 2000 and 2008. However, determining how much aid an institution will offer is important to their budget management as well as their enrollment numbers (Hossler, 2002). According to Gross (2015), the financial aid department has a number of enrollment management-related goals. These may include:

> "maintaining or increasing class size; increasing ethnic diversity; improving academic profile; increasing in net tuition revenue; lowering the tuition discount rate; strengthening weak academic programs; maximizing the return on strong academic programs, and supporting athletic or other specialized programs on campus. (p. 214)"

Hence, it is obvious that institutional financial aid plays a significant role in student recruitment, admission, and retention, regardless of whether it is acting alone or simultaneously with other college departments. Based on the information available, the current study reflected institutional financial aid as one of the significant factors in students' college enrollment decision-making.

Social, Economic, and Policy Context Factors

Finally, changes in social forces, economic conditions, and institutional/public policies can also influence college choice (Perna, 2006). In the current study, two economic conditions (e.g., unemployment rate, inflation rate) are applied in the model as there were no significant events or changes in social forces and institutional/public policies during the time when data was sourced.

The United States experienced several recessions between 1970 and 2009, primarily due to changes in economic policies and government expenditures (Hetzel, 2009; Kotz, 2009). According to the National Bureau of Economic Research (2008), a recession is a prolonged decline in the economy, as indicated by real GDP, real income, employment, industrial production, and wholesale-retail sales. During a recession, individuals experience a decline in their personal income as well as diminished job prospects and, therefore, search for other avenues to increase their earnings. Enrolling in higher education is one of these options. For example, between 1980 and 1992, the value of future earnings differentials between men who graduated high school vs. college increased by 116 percent (Baum, 2001). As well as dwindling labor market opportunities, enrollment in higher education is also heavily influenced by various other factors, such as tuition rates, state aid availability, unemployment rates, and opportunity costs (Betts & McFarland, 1995; Hossler, et al., 1997; Koshal & Koshal, 2000). Therefore, unemployment rate is found to be a significant factor when it comes to determining one's ability to afford higher education (Betts & McFarland, 1995). The decline in labor force demand also tends to reduce the perceived costs of getting a job, particularly for students fresh out of high school. This results in increased college enrollment. Moreover, students forced into postsecondary education by recessions tend to complete the degree and continue their education even when the economy improves (Betts & McFarland, 1995).

There is usually a significant correlation between national unemployment rates and college enrollment (Dellas & Sakellaris, 2003). These findings are consistent across countries, including the United States, where college enrollment is not constrained by a lack of capital. In some countries, a lack of capital overrides rational enrollment decisions. According to Dellas and Sakellaris (2003), human capital theory suggests that participation in higher education should be countercyclical due to opportunity costs, while the ability of an individual to pay for the opportunity appears to be cyclical. Their study examined college enrollment decisions made during four economic downturns from 1968 to 1988. According to their analysis, the propensity to enroll is countercyclical; it shows that every one-percent increase in the unemployment rate, college enrollment increased by .57%. A study conducted by Windolf (1997) also supported to Dellas and Sakellaris (2003) findings showing that there was some modest positive effect of unemployment which generated a short-term increase in college enrollment.

In addition to the national unemployment rate, the inflation rate is also a significant factor impacting college enrollment. In general, inflation involves the rising prices of goods and services and the corresponding decrease in currency value (Truman, 2003). Since inflation impacts college tuition, fees, financial aid, loans, and students' living expenses, both students and colleges are not immune to inflation (Bundick & Pollard, 2019). Garrett (2022) explored higher education enrollment trends along with the inflation rate trend from 1970s to 1980s. He pointed out that undergraduate enrollment trend was flat for the most inflationary periods despite of acute inflation from 1977 to 1982, where inflation rate increased from 6% to 14% (Garrett, 2022). In a related data spanning from 1963 to 2004, Ewing et al. (2010) examined the influence of economic inflation on college enrollment by gender. They observed that in response to an unexpected and sudden increase in inflation, females exhibited an immediate increase in enrollment rates that persisted for two years. Males also responded to the inflation increase, but with a delay and the effect was not as prolonged as it was for females. This indicates that the impact on males is comparatively less severe than on females. Additionally, the study found that female enrollment growth is more enduring than males concerning the impact of inflation on college enrollment. Despite these differences, both males and females seek to accrue human capital through enrollment. Therefore, it can be argued that people tend to mitigate the effects of inflation by accumulating more human capital.

Overall, the high inflation and unemployment rates may persuade more students to invest in their education. This provides generalized insights to higher education administrators that students consider colleges as the only places to invest during a period of economic crisis.

Empirical and Predictive Modeling of College Enrollment

After the number of conceptual college choice models emerged since 1959 (Holland, 1959; Kotler, 1976; Hanson & Litten, 1982; Jackson, 1986; Hossler & Gallagher, 1987; Vossensteyn, 2005; Perna, 2006), many researchers created empirical models for students' college enrollment decisions using influential factors on college choice (Holland & Richards, 1965; Sawiris, 1970; Pickett, 1972; Psacharopoulos, 1973; Kohn et al., 1974). In the late 1980s, a number of researchers implemented statistical regression (i.e., logistic) as a theory to design higher education enrollment models (Bruggink & Gambhir,1996; Cabrera, 1994; Fraysier et al., 2020; St. John & Noell, 1989; Teachman & Polonko, 1988). They conducted empirical models to identify and measure the significance of elements that highly influence students' college choice decisions. Eventually, studies of empirical models for college choice were expanded to the idea of conducting advanced models for predicting students' enrollment (i.e., college choice) decisions (Fraysier et al., 2020). However, as the number of students attending college increased, the types of factors influencing students' college choices became more diverse with abundant data. This led to a significant increase in students' (i.e., applicants') information and made higher education institutions build up vast databases to manage big data (LaValle et al, 2011). The information from these databases provided researchers with a wealth of analytical insights into students' college choice decisions. This allowed educational researchers to employ various machine learning algorithms in addition to LR (Sarker, 2021). Also, it was known that LR algorithm had disadvantages in handling large data for classification prediction, such as leading to an overfitting situation (Dreiseitl & Ohno-Machado, 2002; Rahman et al., 2015). An overfitting situation happens when the model contains too much complexity and features (i.e., attributes) from the large data. This results in a low bias but high variance, which leads LR to make inaccurate predictions (Dreiseitl & Ohno-Machado, 2002). Hence, researchers implemented various machine learning algorithms to overcome the overfitting issue for making predictions, which also handle big data effectively.

Since the early 1990s, machine learning algorithms started to appear as one of the methodologies to identify the major factors that influence students' college enrollment decisions. These algorithms also provided new empirical models to investigate and identify patterns in contemporary higher education admissions (Hossler, 1999; Joseph & Joseph, 2000; LaValle et al, 2011; Bhardwa, 2017). In addition to identifying patterns, machine learning algorithms were implemented to a broader spectrum for predicting students' college enrollment decisions (Delen, 2010; Herzog, 2006; Luan & Zhao, 2006). Hence, the machine learning methods became a vital tool to analyze a large amount of student (i.e., applicant) data, and the predictive analytics using this tool became high demand in the higher education sector (Delen, 2010; Herzog, 2006; Luan & Zhao, 2006). However, compared to other areas of U.S. higher education research (e.g., degree completion, academic performance, and retention), studies on the prediction of "enrollment" using machine learning algorithms are relatively scarce. The following section presents few early studies relevant to developing prediction models for U.S. college enrollment using machine learning algorithms. Since the present study proposes to use four different machine learning algorithms, the next two sections explain the concept of machine learning for predictive modeling in general and why the current study came up with the idea of using four machine learning algorithm models.

Concept of Machine Learning

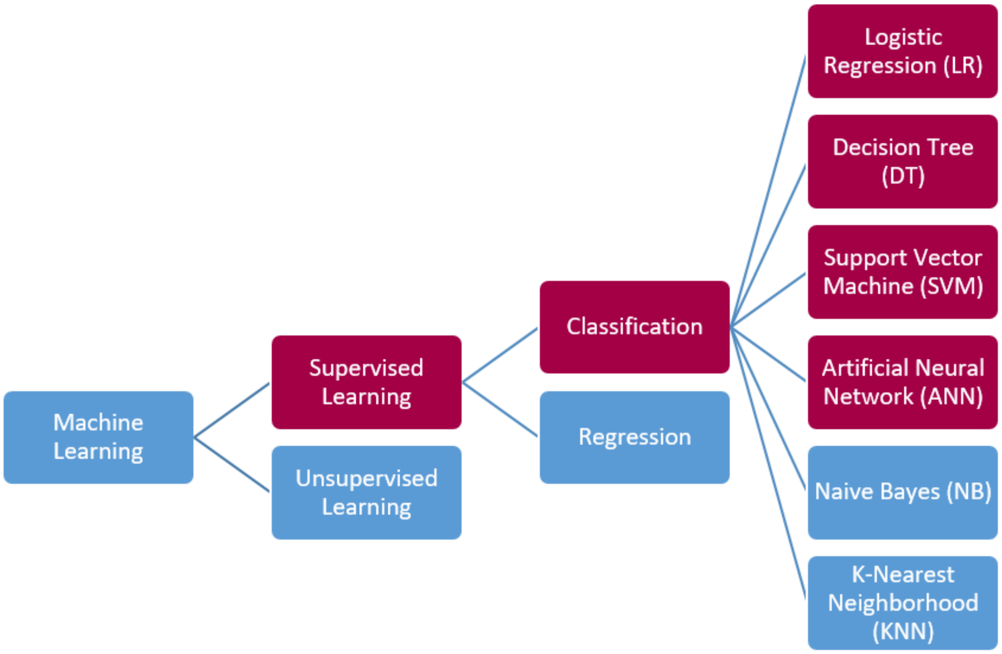
Machine learning combines statistics, mathematics, and computer science into one problem-solving pathway (Dara, et al., 2022). This involves discovering patterns, training, and testing data to create computer programs that automatically recognize complex structures and make intelligent decisions (Michell, 1997). There are several different types of machine learning algorithm approaches – supervised learning and unsupervised learning (Sharma & Kumar, 2017).

A supervised learning approach uses a set of labeled input data and corresponding output data. It trains a model to map labeled inputs to outputs so it can predict the outputs to any new set of input data. All supervised learning-related algorithmic methodologies take the form of either classification or regression. Classification methodologies predict discrete responses as outputs, which can be classified into two different groups. Regression methodologies, on the other hand, predict continuous responses. In supervised learning, the goal is to predict outcomes for new data and these outcomes are known up front the type of results to expect. Unsupervised learning approach uses unlabeled datasets that do not contain explicit instruction on what to do with it. Hence, unlabeled datasets used here only have input data and no corresponding output variables. The goal of unsupervised learning is to discover hidden patterns and underlying structures/distributions of input data in order to learn more about data. For the purpose of the present study, I am interested in predictive modeling from a classification perspective which predicts output into two different groups (i.e., enrolled vs. non-enroll) using labeled data. Hence, a supervised learning approach is used.

As mentioned earlier, machine learning models under the supervised learning approach can perform classification models using various algorithms (Alpaydin, 2011). However, it is crucial to select the appropriate tool from the machine learning toolbox for a given data set.

For the current study, certain machine learning algorithmic methodologies/models are chosen based on the following criteria. Research and experimentation that are currently taking place but have not been used for comparative analysis. The description of past studies using machine learning for U.S. College enrollment prediction is stated in the following section for 'Implementation of Machine Learning Algorithms for College Enrollment Prediction Modeling.' Based on these criteria, four primary algorithms are selected to compile machine learning models: ANN, SVM, DT, and LR. In addition, Python 3.10. software is used for analysis. Figure 2 shows a simplified classification diagram from machine learning algorithms including LR, DT, SVM, and ANN.

Figure 2. Classification of Machine Learning Algorithms including LR, DT, SVM, and ANN



Logistic Regression (LR)

Logistic Regression (LR) is a supervised classification algorithm in machine learning that predicts the outcome of a categorical dependent variable by using the probabilities of achieving the output categories (Green & Salkind, 2014; Warner, 2013). The dependent variable in LR is categorical, and the algorithm uses the available data to construct a model that calculates the unknown outcome variable, similar to linear regression (Warner, 2013). In LR, input values x are linearly combined through weights or coefficients to predict a real-valued output y, but unlike linear regression, LR estimates the probability of the dependent variable *y* belonging to each class (Warner, 2013). The general formula for this learning technique with a single independent variable *x* and a dependent output variable *y* is shown in Equation (1).

$$\mathrm{P}(y = 1) = \sigma(\beta_0 + \beta_1 x) \qquad (1)$$

During training, Logistic Regression (LR) learns the bias coefficient β0 and the coefficients β1 for each independent input value *x*. LR uses a mathematical function called the sigmoid function, denoted by σ, to map the linear combination of inputs into the range of 0 to 1, which provides probabilities for each output category. The sigmoid function is defined by Equation (2). When multiple independent input variables are involved, the input values xs are represented as a vector *x* with a corresponding set of coefficients to be learned.

$$\sigma(z) = \frac{1}{1+e^{-z}} \qquad (2)$$

As an illustrative example, consider the problem of identifying if a student will enroll in a college based on the amount of institutional scholarships. In this case, we can define the input independent variable as amount of scholarships which take a numerical value and the output variable as student's decision to enroll which takes values 0 (i.e., not enroll) or 1 (i.e., enroll). Then our LR problem will represent the probability of a student to enroll in college given amount of institutional scholarships and is shown in equation (3). In LR, a general rule is that if the probability is greater than 0.5, the decision is considered true (i.e., 1), otherwise, it is false (i.e., 0). Therefore, the LR model predicts whether the student will enroll in college by calculating the probability and thresholding it based on 0.5, as demonstrated in equation (3).

$$\mathrm{P}(\mathrm{y} = 1) = \mathrm{P}(Student_{Enroll} = True)$$

$$= 0.82 * (Institutional\ Scholarship) - 0.32 \qquad (3)$$

Using LR models to describe research findings offers some advantages based on the LR algorithm theory. Compared to other ML algorithms, building LR models is less computationally complex and requires less time for training computation (Tu, 1996; Ayer, 2010). Also, since LR models are conducted based on statistical methods, the predicted probability of the outcome can be easily calculated (Dreiseitl & Ohno-Machado, 2002; Mehta & Patel, 1995). Moreover, most statistical software packages utilized for constructing LR models provide confidence intervals, probability of outcome, and standard output (Dreiseitl & Ohno-Machado, 2002). Consequently, LR models can easily identify the most predictive variables of an outcome by examining the co-efficients and the corresponding odd ratios (Dreiseitl & Ohno-Machado, 2002; Tu, 1996).

Despite some advantages, LR also has some drawbacks. Constructing LR models is more challenging than other ML algorithms because it requires expert domain knowledge, including an understanding of statistical concepts such as multicollinearity (Harrell et al., 1996). Additionally, LR models can only incorporate complex relationships of input variables if they are explicitly identified as affecting outcome variables (Ranganathan et al., 2017). This suggests that LR models may be susceptible to overfitting in high dimensions, as they may involve complex relationships among predictor (i.e., input) variables and outcome variables (Dreiseitl & Ohno-Machado, 2002).

Decision Trees (DT)

Decision Tree (DT) is a supervised machine learning algorithm that is useful for solving classification problems. It is suitable for dealing with continuous, categorical, and binary input and output variables (Liu et al., 2017). In situations where the output variable is categorical, decision trees are also known as classification trees (Liu et al., 2017). The fundamental idea behind a decision tree is to employ a tree-like flowchart structure to make predictions, where each branch represents a choice between different attribute options in the internal nodes, which ultimately leads to a final decision in the leaf node (Yang, 2019).

The typical method for constructing a decision tree from the training data involves partitioning the entire dataset at the root node into subsets using a specific criterion, usually a decision about a feature (Wang et al., 2005). This process of splitting based on different internal node features continues until either a subset at a node has the same values as the target variable, or further splitting no longer improves predictions (Yang, 2019). The primary goal of decision trees is to determine the optimal split for each node of the tree. However, evaluating the quality of a particular split is often a matter of subjectivity.

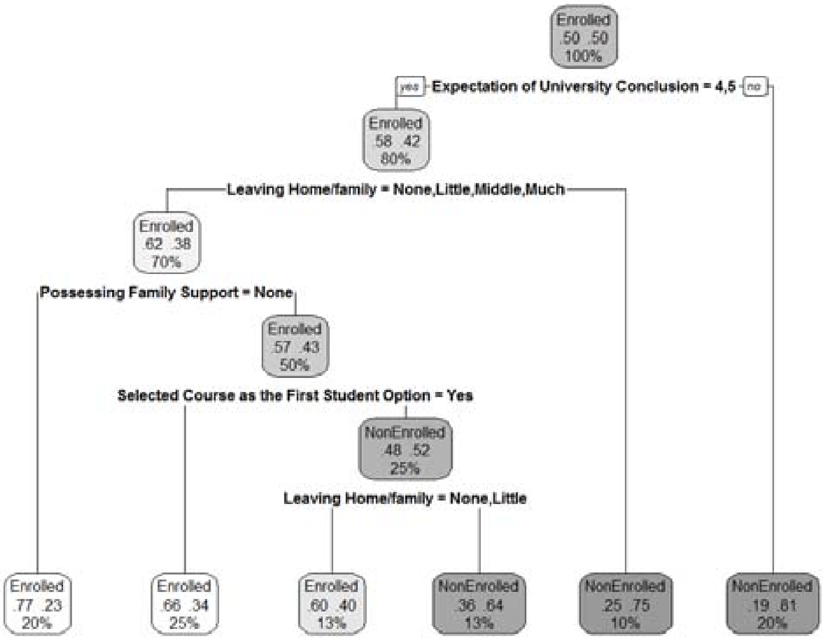
When evaluating the quality of splits in DT, various metrics are used. The two primary metrics used to evaluate splits in decision trees are Gini Impurity and Information Gain (Raileanu & Stoffel, 2004). Gini Impurity measures the impurity of a partitioned dataset and determines how often a randomly chosen element from the set is mislabeled based on the distribution of the labels in the subset. This metric is at its minimum (i.e., zero) when all cases in the node belong to a single target category. In contrast, Information Gain is used to compute the expected quantity of information required to determine whether a new instance should be classified as a yes or no for each node of the tree, given that the example has reached that node. The node with the highest Information Gain value is considered to have the best split.

For illustration purposes, consider the DT shown in Figure 3 which describes the probabilities for the student enrollment decision (Gomes & Almeida, 2017). The tree is constructed using Gini Impurity as the split evaluation metric. From the tree, it is possible to make the following inferences:

> "… if students answer option 4 or 5 (high expectation) concerning their "Expectation of University Conclusion", and if they perceive "No, Little, Middle, or High" difficulty related to "Leaving Home/Family", and if they see "Little, Middle, High or Very High" difficulty related to "Possessing Family Support" and if they answer "No" for

> the “Selected Course as the student’s first option” and if they answer “None or Little” difficulty for “Leaving Home/Family”, there is a 60% likelihood that these students are enrolled (p.8).”

Figure 3. The Generated Tree from the Decision Tree Model: Classification of the College Enrollment Decision (Gomes & Almeida, 2017)



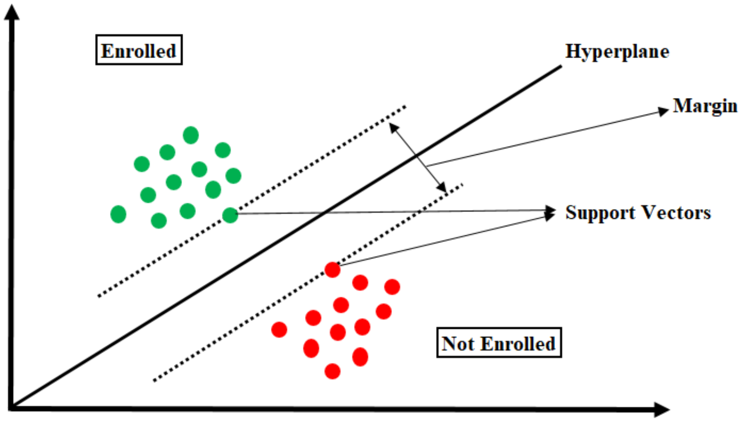
Note. Adapted from “Advocating the broad use of the decision tree method in education (p. 7),” by C.M.A., Gomes & L. S. Almeida, 2017.

Utilizing the DT algorithm to present research outcomes offers some benefits. DT can simplify intricate relationships between input and target variables by dividing the original input variables into meaningful subgroups (Statnikov et al., 2008; Song et al., 2015). Additionally, DT is straightforward to comprehend and interpret, providing a non-parametric approach that does not rely on distributional assumptions (Song et al., 2015). However, the DT algorithm also has certain drawbacks. The primary disadvantage is that it is prone to over- and underfitting, particularly when working with small datasets, which can restrict the models' generalizability and robustness (Statnikov et al., 2008). Another potential issue is that the DT algorithm depends on the order of the attributes/variables, which can influence the prediction outcome. Consequently, one should exercise caution when interpreting DT models and their results (Jijo & Abdulazeez, 2021). Finally, DT necessitates a lengthy training time, which may not be suitable for efficiency when working with large datasets (Jadhav & Channe, 2016).

Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a classification algorithm suitable for various applications (Parikh & Shah, 2016). SVM represents each data point, consisting of n features, as a point in an n-dimensional space, where each coordinate denotes the value of a specific feature. To classify the data, the algorithm identifies the optimal hyperplane that can segregate the two classes. A hyperplane is a line that divides the space of input features. For instance, in a binary classification task where a student's college enrollment decision depends on a single influential factor x, and the output variable y denotes the enrollment decision, the problem can be visualized in two-dimensional space, as illustrated in Figure 4. The SVM classifier locates a hyperplane that separates all the input points. Once obtained, the line is used to classify the data points by inputting their corresponding values. If the equation returns a positive value, the point is classified as belonging to the first class, whereas a negative value indicates that the point belongs to the second class. Points close to the line have values close to zero and may be difficult to classify. The margin is the perpendicular distance between the line and the closest data points.

Figure 4. Example of SVM Classification for College Enrollment Decision



The optimal hyperplane for class separation is determined by the largest margin between the two classes, with the closest points called support vectors. These vectors help define the line and build the classifier. The hyperplane is learned through an optimization process that maximizes the margin on the training data. In real-world scenarios, data is typically noisy and cannot be perfectly separated by a hyperplane. In these cases, the SVM will relax the maximum-margin hyperplane constraint by allowing some training data points to violate the separation principle. The amount of violation allowed by the classifier is defined by a tuning parameter, referred to as the ‘C’ (i.e., Cost parameter) (Hastie et al., 2004). Generally, different C values are tried, and the one that best fits the data is selected.

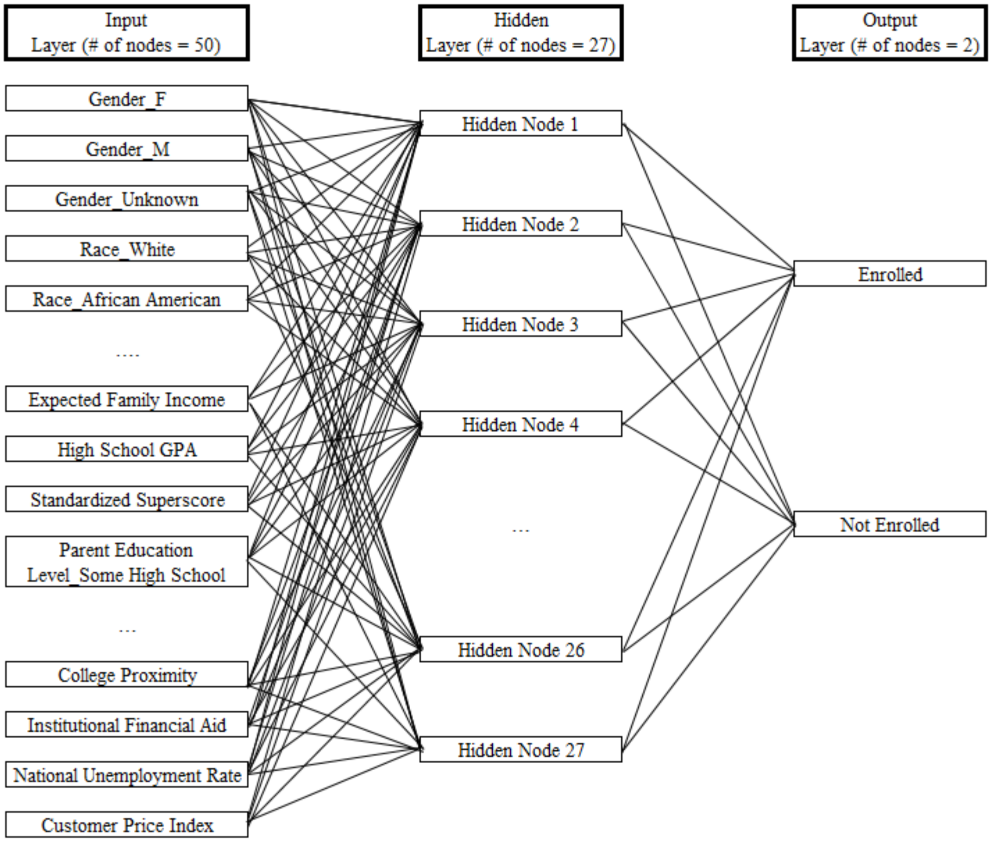
SVM is capable of creating a linear hyperplane between two classes with ease. However, when it comes to classifying non-linear data, SVM applies the kernel trick technique. The kernel trick involves the use of functions that convert a low-dimensional, non-separable input space into a higher-dimensional, separable space. The commonly employed kernels are linear, polynomial, and radial basis function.

Conducted models based on the SVM algorithms describing research findings offer some advantages. SVM works well on a dataset with many features (i.e., predictive variables) (Auria & Moro, 2008). Also, it provides a clear margin of separation, which effectively works for classification. Similar to ANN, SVM gives good classification results even if there is not enough information about the data, including unstructured data. This also implies that SVM can solve complex classification problems using a convenient kernel solutions function (Auria & Moro, 2008; Fedorovici & Dragan, 2011). However, SVM also has some disadvantages. SVM requires a very long training time; hence, implementing large datasets is not highly recommended (Yu et al., 2004). Also, as mentioned earlier that SVM is good at solving complex classifications using kernels, but it is challenging to choose the appropriate kernel for the solution (Auria & Moro, 2008; Fedorovici & Dragan, 2011).

Artificial Neural Networks (ANN)

Artificial Neural Network (ANN) is a supervised machine learning algorithm suitable for both classification and regression prediction tasks (Lau et al., 2019). It is made up of nodes arranged in three layers - input, hidden, and output layers - with each layer having several neurons. The input layer has input nodes that represent input variables, which are predictors of the outcome, while the output layer has output nodes that represent the predicted outcome (e.g., students' enrollment decision). The intermediate values calculated by the network, which do not have any meaning, are stored in the hidden layer's nodes. The hidden nodes enable the ANN to model complex relationships between the input variables and the outcome. An optimal number of hidden layer(s) and neuron(s) can be chosen based on common rules of thumb, such as setting the number of hidden layers equal to one and the number of neurons in that layer equal to the mean of the neurons in the input and output layers (Thomas et al., 2017). For illustration, Figure 5 shows the ANN with 50 input nodes (i.e., 50 influential metrics on college enrollment decisions), 27 hidden nodes, and two output nodes (i.e., enrolled, non-enrolled).

Figure 5. Example of ANN classification architecture of College Enrollment Decision



In an Artificial Neural Network (ANN), the connections between nodes in different layers are established by means of connection weights that represent the strength of the relationship between variables, similar to the coefficients in a logistic regression model. The ANN "learns" the relationship between input variables and outcome by adjusting the values of these connection weights based on known cases. This process of estimating the optimal weights that generate the most reliable outcomes is called learning or training, which is analogous to estimating parameters in logistic regression. However, an ANN is not simply an automated logistic regression model because they use different training algorithms for parameter estimation. There are several algorithms for training ANNs, including forward and backward propagation. During forward propagation, the ANN computes the predicted output for each instance in the data set, compares it with the actual output, and calculates the error between them. In contrast, backward propagation involves adjusting the connection weights associated with each input using gradient descent to minimize the discrepancy between actual and predicted outcomes by propagating the disparity from the output node to the input nodes. Overall, the backward propagation training algorithm is the most widely used method for training ANNs.

According to the theory behind the ANN algorithm, implementing an ANN model has certain advantages. One such advantage is that building an ANN model requires less domain knowledge compared to other ML models like LR. This is because there are many user-friendly software interfaces available that can quickly build ANN models without requiring an in-depth understanding of the network's structure (Harrell et al., 1996). Furthermore, ANN is well-suited for modeling without considering multicollinearity because it does not require any prior knowledge about the data underlying the model. ANN can automatically detect and model any arbitrary relationships between input and output variables (Hansen & Sargent, 2001; Dreiseitl & Ohno-Machado, 2002; Bejou et al., 1996; Tu, 1996). In addition, ANN can model any implicit interactions among input variables. Detecting interactions among input variables is often difficult, but ANN can handle these complex interactions by using hidden nodes, which act as interaction detectors and increase the network's capacity to learn complex relationships among the predictor variables (Ayer et al., 2010).

The ANN algorithm also comes with certain drawbacks, similar to other ML algorithms. One of the main disadvantages is the increased risk of overfitting due to its complex structure. Large networks with more hidden nodes are particularly susceptible to overfitting because they tend to detect almost any possible interaction, making the model too specific to the training dataset. (Hansen & Sargent, 2001; Ranganathan et al., 2017; Ayer et al., 2010). Unlike LR, ANN models are not primarily designed for statistical use, which makes it difficult to generate confidence intervals of the predicted probabilities. This often requires extensive computations (Ranganathan et al., 2017).

Model Evaluation and Comparison

Once each model is trained using the training dataset, I implemented the test dataset to these models to measure the classification prediction performance. In this process, a confusion matrix was generated for each ML algorithm model. The confusion matrix was used to calculate the model classification accuracy, sensitivity, specificity, precision, F_1 scores, and plot the ROC and PR curves for the models. Based on the ROC and PR curves, the AUC scores were calculated to verify if the model results were unbiased with respect to imbalanced data. If the cross-validated AUC score is similar to the one which is determined from the test data, it means that the model is actually learning from the training data. All the above-mentioned performance metrics were used to compare the models to find the one that best suits for prediction of student enrollment decisions. Moreover, it gave a better idea of generalizing the model's prediction performance based on the new data, which were never seen previously.

Implementation of Machine Learning Algorithms for College Enrollment Prediction Modeling

As mentioned earlier, machine learning algorithms appeared in the early 1990s in the realm of higher education when Song and Chissom (1993) proposed using a neural network algorithm to create a U.S. college enrollment prediction model. Their research used a comparative analysis of the enrollment models using an artificial neural network (ANN) versus time series analysis. Later in that decade, Shah and Sastry (1999) developed a college enrollment prediction model using a decision tree (DT) algorithm. Since LR often performs poorly in classification problems, Shah and Sastry (1999) constructed several binary DT algorithms, as originally proposed by Friedman et al. (1996). Other few studies related to college enrollment prediction modeling used two machine learning algorithms (ANN, DT) along with LR for either a single-case or comparative analysis (Walczk & Sincich, 1999; Gonzalez & DesJardins, 2002; Antons & Maltz, 2006; Chang (2006); Bruggink & Gambhir, 1996; Cabrera, 1994; Fraysier et al., 2020; Breiman, 2001). However, no additional studies using new machine learning algorithms appeared through 2010.

In 2011, a group of scholars published a comparative analysis proposing two new machine learning algorithms (Vialardi et al., 2011). The study implemented DT, K-Nearest Neighbors (KNN), and Naïve Bayes (NB) algorithms for college enrollment prediction and model comparison purposes. The results showed that the DT provided better classification accuracies than KNN and NB in terms of predicting students’ enrollment decisions. In addition to the appearance of the KNN and NB algorithms, support vector machine (SVM) and random forest (RF) algorithms have appeared in higher education enrollment prediction modeling research since

2014 (Ragab et al., 2014; Cirelli et al., 2018; Slim et al., 2019; Lux et al., 2016). Since various types of machine learning algorithms have been used for college enrollment prediction modeling, several comparative studies have used these algorithms to determine which one is the most accurate in terms of classification. Table 2 presents a summary of U.S. college enrollment prediction studies conducting comparative analysis using different types of machine learning algorithms.

Table 2. Summary of U.S. College Enrollment Studies Using Machine Learning Algorithms for Comparative Analysis

College Enrollment Prediction Modeling	Types of Machine Learning Algorithms Used	Best Prediction Model
Walczk & Sincich (1999) DesJardins & Gonzales (2002) Gerasimovic & Bugaric (2018)	LR vs. ANN	ANN
Antons & Maltz (2006) Chang (2006)	LR vs. ANN vs. DT	LR ANN
Vialardi et al. (2011)	DT vs. KNN vs. NB	DT
Ragab et al. (2014)	ANN vs. DT vs. SVM vs. KNN vs. RF	DT
Lux et al. (2016)	ANN vs. SVM	ANN
Cirelli et al. (2018)	LR vs. ANN vs. NB vs. RF vs. SVM	SVM
Slim et al. (2019)	LR vs. SVM	SVM
Basu et al. (2019)	LR vs. NB vs. DT vs. SVM vs. KNN vs. RF vs. GB	LR

Table 2 shows that the best model for predicting college enrollment decisions varies across studies. Among comparative analysis studies of multiple ML algorithms of conducting college enrollment prediction models, some studies concluded that ANN is the better model with the highest accuracy rate (Walczk & Sincich, 1999; DesJardins & Gonzales, 2002; Gerasimovic & Bugaric, 2018; Chang, 2006; Lux et al., 2016). In addition to the high accuracy rate, ANN was identified to be the best algorithmic methodology since it adapts easily to related independent variables without the appearance of a multicollinearity problem (Gerasimovic & Bugaric, 2018).

Moreover, ANN can recognize the appearance of nonlinearity and interactions in input data in contrast to LR (Chang, 2006; Lux et al., 2016).

On the other hand, studies concluded SVM as a best algorithm stated that this model has the highest accuracy and easily implementable to the big data. These studies pointed out the advantage of SVM that it is usually used with data that have a large number of predictor variables (i.e., input variables; college enrollment decision factors) (Cirelli et al., 2018; Slim et al., 2019). Other studies concluded that DT (Vialardi et al., 2011; Ragab et al., 2014) or LR (Antons & Maltz, 2006, Basu et al., 2019) showed highest accuracy with low prediction errors. Hence those are the best algorithms to compile a predictive model for college enrollment.

Overall, the preferable machine learning algorithms for college enrollment prediction are ANN, DT, SVM, and LR, which were chosen to be the best models in the past studies. However, no comparative analysis has been conducted using all four algorithms. Hence, the purpose of the present study is to create college enrollment prediction models using these four methods and then conduct a comparative analysis to see which model is the best fit for predicting college enrollment decisions among students in the next academic cycle. This study presents an exploration and comparative study of predictive analysis using three machine learning algorithms – ANN, SVM, DT, and LR in the area of university student intake. The conceptual overview of four algorithms is described in the methodology section of chapter three.

Summary

This chapter introduces and explains the background of college choice (i.e., enrollment decision) and related models. Among various college choice models, the present study proposes using the most recent college choice model created by Perna (2006). Her model is a conceptual framework for segmenting the attributes that influence students' enrollment decisions (Perna, 2006).

Once the college choice model is selected, eleven selected attributes that influence students' college enrollment decisions are reviewed based on Perna's model. This section explains the historical influences of those factors on students' college enrollment decisions. The current study also proposes using all eleven attributes to create the college enrollment predictive modeling using machine learning.

In addition to illustrating the theoretical college choice model and related factors, the implementation of the machine learning methodologies on college enrollment modeling is presented. But first, it covers the concept of machine learning and how it is applied for predictive modeling in general. It provides basic knowledge of machine learning, including how the four different machine learning algorithmic methodologies are derived.

After going over the general concept of machine learning algorithms, it covers past studies using machine learning methodologies for conducting U.S. college enrollment predictive models. Based on the past research, it was found that no comparative analyses were conducted for specific machine learning algorithms (i.e., the four algorithms) which were identified to be the best from other research. Hence the present study, once again, proposes to use four different machine learning algorithms to conduct college predictive modeling and comparative analysis to see which model is the best fit for prediction. The next methodology chapter covers how these four different machine learning methodologies are used for predictive modeling along with data collection and design.

The enrollment management process is directly involved in new student recruitment, financial aid planning and budgeting. In order to maximize student enrollment, enrollment management administrators should frequently communicate with the admissions office and financial aid department to set up better plans on student recruitment along with allocating institutional scholarships and grants. Considering the large expenditure of institutional marketing and financial aid awards, this research study explores different attributes that influence college choice (i.e., enrollment) of admitted students and implements machine-learning algorithm models for predicting admitted students' enrollment decisions.

The objectives of the current study are 1) to provide a better understanding of both student and institution-related attribute(s) that highly influence students' enrollment decisions and 2) to identify better predictive model(s) using machine learning algorithms that institutions can discern admitted students who are highly likely to enroll relative to others. This allows enrollment management to support financial aid, admission, and marketing departments to assign budgets effectively and thus not only maximize enrollment but also save a big portion of marketing efforts efficiently for recruiting students. In this chapter, I discuss the factors and enrollment prediction models using machine learning algorithms, including a wide range of features related to student and institutional characteristics.

CHAPTER THREE

METHODOLOGY

This chapter describes the data collection, data transformation, dataset split with the k-fold cross-validation method, machine learning algorithms methods for prediction modeling, and comparative analysis methods of college enrollment prediction models. The research questions that guid the analysis of the study include:

1. Which machine learning prediction model (Logistic Regression vs. Artificial Neural Network vs. Decision Tree vs. Support Vector Machine) demonstrates the best prediction performance for testing Fall 2019 enrollment after training the historical data (from Fall 2013 to Fall 2018)?
2. With a machine learning prediction model chosen in Q1, what are the most significant factors contributing to the prediction of the enrolled students for Fall 2019?
3. With a machine learning prediction model chosen in Q1, what applications can be followed up and utilized to support a higher education institution's enrollment management strategy as a plan of action?

Data Description

The data for this study was derived from two sources. One is from the institutional database of one of the 4-year private Jesuit universities in the mid-west urban area. It was housed in the Student Data Management Warehouse called Slate. All students' application data were collected by the university and stored in Slate when they submit the application. The application data include attributes of students' habitus (i.e., gender, socioeconomic status, race, academic performance), their high school and community environment (i.e., parental education level, secondary school academic climate), university characteristics which students applied (i.e., proximity, institutional financial aid), and social economic and policy context (i.e., national unemployment and inflation rates). These data were required for students to provide in their applications. Hence there were no missing data in the dataset. In addition, rates of national unemployment and inflation data were obtained from the Bureau of Labor Statistics (BLS) database, which comes from another source besides Slate. The study was also submitted to the college's Institutional Review Board (IRB) and was determined to be exempt in December 2022 (see Appendix C).

The de-identified students' application data from Slate were exported into a .csv file to be analyzed. The data consisted of 53,189 of admitted students and 11,336 of those were identified as enrolled as a subset of the admitted student pool. These students' records were from six cohort years starting from 2013 to 2019. The data consisted of a variety of information about each student, such as cohort year, gender, race, high school academic performances (ACT/SAT scores and High School GPA), Parental education level, Expected Family Contribution (EFC), High School academic climate, institutional financial aid amount, and proximity from their residence to college. Also, the study used data from the BLS for collecting the 'Civilian Unemployment Rate' and 'Consumer Price Index (CPI)' from 2013 to 2019. Hence, two datasets from Slate and BLS were merged for data preparation. Overall, a total of 12 attributes were captured for each student including the output variable (i.e., enrollment decision). Table 3 describes the features and their data types. The following sections 'Data Transformation for Categorical Variables' and

‘Data Description of Numerical Variables,’ describe the detailed steps taken to transform the data of certain attributes for conducting analysis.

Table 3. Description of Data Fields for Admitted Students from Slate and BLS

Data Source	Attribute No.	Perna's (2006) College Choice Theoretical layer	Attributes	Data Type	Description
Student Data Management Warehouse (i.e., Slate)	1	-	Cohort Year	Categorical	2013, 2014, 2015, 2016, 2017, 2018, 2019
	2	Student's Habitus (Layer 1)	Race	Categorical	White
					Hispanic or any race
					Black or African-American
					Asian
					Other
					Unknown
	3		Gender	Categorical	Male, Female, Unknown
	4		High School GPA	Numerical	Range from 0.00 to 4.00
	5		Standardize Test Superscore	Numerical	Range from 0 to 36
	6		Expected Famliy Contribution	Numerical	Range from $0 to $999,999
	7	Secondary School and Community Context (Layer 2)	Parental Education Level	Categorical	Some High School
					High School Graduate
					Associates Degree
					Some College
					Bachelors Degree
					Graduate Degree
	8		High School Academic Climate	Categorical	51. Public schools primarily serving traditional, blue-collar populations
					52. Private/religious schools primarily serving Puerto Rican/Caribbean/ESL populations
					53. Comprehensive public/religious schools primarily serving traditional, blue-collar communities
					54. Public schools in rural settings primarily serving African American and Hispanic populations
					55. Private/religious schools predominantly serving males from racially diverse populations
					56. Public/private schools serving racially diverse populations with a strong interest in athletics
					57. Public schools in urban settings primarily serving African American populations
					58. Public/private schools primarily serving Jewish populations
					59. Public schools in suburban settings primarily serving white, blue-collar populations

Table 3. Description of Data Fields for Admitted Students from Slate and BLS (Continue.)

Data Source	Attribute No.	Perna's (2006) College Choice Theoretical layer	Attributes	Data Type	Description
Student Data Management Warehouse (i.e., Slate)	8	Secondary School and Community Context (Layer 2)	High School Academic Climate	Categorical	60. Private schools primarily serving Jewish female populations
					61. Private schools in urban settings serving racially diverse populations
					62. Public schools serving Hispanic populations with traditional values
					63. Public schools in urban settings primarily serving Hispanic/ESL and African American populations
					64. Public schools primarily serving Asian/ESL populations
					65. Public schools in suburban settings serving affluent, racially diverse populations
					66. Public/private schools primarily serving women from racially diverse populations
					67. Religious/private schools primarily serving women, upper-middle-class populations
					68. Religious schools primarily serving Catholic populations
					69. Public schools primarily serving African American populations
					70. Public schools primarily serving affluent suburban populations
					71. Public/private/religious schools primarily serving Puerto Rican/Caribbean/ESL populations
					72. Homeschoolers and private/religious schools primarily serving upper-middle-class Christian populations
					73. Public schools in urban settings primarily serving Hispanic (particularly Mexican) populations
					74. Private schools primarily serving Asian/ESL populations
					75. Public schools in rural settings primarily serving middle-class populations with traditional values
					76. Private schools primarily serving affluent, racially diverse populations
					77. Private schools in urban settings serving racially diverse populations
					78. Public schools in small town and suburban settings serving vocationally diverse populations
					79. Public schools primarily serving highly educated, middle-class populations
	9	Higher Education Context (Layer 3)	Insittuion Financial Aid	Numerical	Range from $0 to $76,762
	10		Proximity		Range from 1 mile to 9,361 miles
	11	-	Enrollment Decision	Categorical	Not Enrolled, Enrolled
Bureau of Labor Statistics	12	Social economic, and Policty Context (Layer 4)	National Unemployment Rate	Numerical	Range from 3.6% to 7.6%
	13		National Inflation Rate		Range from 1.8% to 2.2%

Data Transformation for Categorical Variables

Data transformation involves converting raw data into a usable dataset for analysis. To prepare the data for input into machine learning algorithms, certain data transformation techniques were applied to convert some data variables into different data types. This step is crucial because the dataset contains many categorical features that must be converted to numerical values before any further analysis can be performed.

To convert the categorical data variables into a usable format for machine learning algorithms, a popular technique called one-hot encoding was employed (Seger, 2018; Okada et al., 2021; Rodriguez et al., 2015; Cerda et al., 2018). This technique is essential in improving the prediction and classification accuracy of a model. One-hot encoding involves creating a new binary feature for each possible category, with a value of 1 assigned to the feature of each sample that corresponds to its original category, and a value of 0 assigned otherwise. The dataset contained six categorical metrics (Race, Sex, Parental Education Level, High School Academic Climate, and Enrollment Decision), which were one-hot encoded, resulting in 45 metrics. In the following subsections, each metric is described in detail, along with how it was transformed for use in the study.

Race

Student race is a categorical variable which was collected based on students' application responses. This study employed 11 racial categories that conformed by the Integrated Postsecondary Education Data System (IPEDS) classification system: White, Black or African-American, Hispanic of any race, Asian, Nonresident Alien, American Indian or Alaska Native, Native Hawaiian or other Pacific Islander, Two or More Races, Other, Multiple Ethnicities or unknown, and Race/Ethnicity Unknown. Due to the significantly small number of records classified under Nonresident Alien, American Indian or Alaska Native, Native Hawaiian or other Pacific Islander, Two or More Races, Other, Multiple Ethnicities or unknown, and Race/Ethnicity Unknown, these categories are grouped into two different groups. In this study, I classified Nonresident Alien, American Indian or Alaska Native, Native Hawaiian or other Pacific Islander, Two or More Races, Other under the name of "Other" and Multiple Ethnicities or Unknown, and Race/Ethnicity Unknown as "Unknown." The other categories of White, Black or African-American, Hispanic of any race, Nonresident Alien, Asian remained separate. Hence, the total number of categories of race is reduced to six. Based on the re-categorization, a series of seven dummy variables were then created including White, Black or African-American, Hispanic or any race, Asian, and Nonresident Alien, Other, and Unknown. Table 4 shows an example of how race is coded through a one-hot coding data transformation process.

Table 4. One-Hot Encoded for Race

White	Black African-American	Hiapnic or any race	Asian	Other	Unknown
1	0	0	0	0	0
0	0	0	0	0	1
0	0	1	0	0	0
0	0	0	1	0	0
0	0	0	0	1	0

Sex

Similar to race, student sex was a categorical variable which was collected based on students' application responses, as well. There were two different categories under sex: Male and Female. Since sex has two different categories, label coding was considered at first as coding Female as 1 and male as 0. However, the label coding makes it seems that there is a ranking between values. Hence, one-hot encoding was applied with creating a series of two dummy variables as part of data set consisting of Female and Male. Table 5 shows an example of how sex was coded through a one-hot coding data transformation process.

Table 5. One-Hot Encoded for Sex

Female	Male
1	0
0	0
0	1
1	0

Parent Education Level (PEL)

PEL was a categorical variable which was also collected based on students' application responses. There were 6 categories under PEL: Some High School, High School Graduate, Associate Degree, Some College, Bachelor's Degree, and Graduate Degree. Since PEL has six different categories, a series of 6 dummy variables was created as part of data set consisting of Some High School, High School Graduate, Associate Degree, Some College, Bachelors Degree, and Graduate Degree. Table 6 shows an example of how PEL is coded through a one-hot coding data transformation process.

Table 6. One-Hot Encoded for Parent Education Level

Some High School	High School Graduate	Associatees Degree	Some Degree	Bacehlors Degree	Graduate Degree
1	0	0	0	0	0
0	0	1	0	0	0
0	1	0	0	0	0
0	0	0	0	0	1
0	0	0	0	1	0
0	0	0	0	0	0
0	0	0	1	0	0

High School Academic Climate (HSAC)

As it was described in Table 8, HSAC was also a categorical variable involving 29 different categories. HSAC data was collected based on the students' high school information that were provided for application response. Once students' high school information was collected, these information were sent out to the College Board to get a High School Cluster tagging service which segments all high schools into 29 descriptive clusters (segment labeling starts from 51 to 79). Generated clusters by the College Board represent more than 33,000 high schools in the U.S. They were defined by 40 academic and demographic factors and described by 51 characteristics that influence students' college choice decisions. Once the College Board completed the High School Tagging service by matching 29 descriptive high school clusters for each student record, this data were imported to Slate to merge with student application data. See Appendix B for detailed information about each high school segmented cluster. As mentioned above, there were 29 different categories under HSAC. Hence, a series of 29 dummy variables was created as part of data set consisting for all 29 categories. Table 7 shows an example of how HSAC was coded through a one-hot coding data transformation process.

Table 7. One-Hot Encoded for High School Academic Climate

HSAC_51	HSAC_52	HSAC_53	…	HSAC_77	HSAC_78	HSAC_79
1	0	0	…	0	0	0
0	0	1	…	0	0	0
0	0	0	…	0	1	0
0	0	0	…	1	0	0
0	1	0	…	0	0	0
0	0	0	…	0	0	1
0	0	1	…	0	0	0

Enrollment Decision

Students' enrollment decision data was automatically stored in Slate once students made their enrollment decision. In general, such data was coded and. categorized as 'Enrolled' vs. 'Not Enrolled'. Like one-hot encoding schema, this metric was converted into numerical data using binary coding. For example, students who enrolled are coded as 1, whereas students who did not, are coded as 0. This metric was a dependent binary outcome variable for all four machine learning models (i.e., Artificial Neural Network, Decision Tree, Support Vector Machine, Logistic Regression).

Data Description of Numerical Variables

In addition to the five categorical metrics described earlier, another seven metrics have been identified as numerical and continuous. This section described numerical metrics and how those are defined and stored in the databases.

High School GPA (HSGPA)

Students must submit their HSGPA to get their application evaluated for admission. Since the way of high school GPA calculations varies across highs schools, the institution decided to recalculate all the students' GPAs under a 4.0 scale for consistency. Hence, range of HSGPA was from 0.0 to 4.0.

Standardized Test Superscore

Students were also required to submit their standardized test scores such as ACT or SAT along with their HSGPA. These scores were usually reported directly from ACT or the College Board under a superscore scale. Superscores were eligible for students who have taken the standardized tests more than once. A superscore was the average of a student's best scores from each subject from multiple test attempts. Since more than 80% of students reported their standardized scores on ACT scale, students with SAT superscores were converted into the ACT scale based on the ACT and College Board's converting table (ACT, 2018). Hence the Standardized Test Superscore metric ranged from 0 to 36.

Estimated Family Contribution (EFC)

As mentioned earlier, EFC was a continuous numerical variable. Hence EFC variable was not necessary to go through the one-hot coding for creating dummy variable(s). This variable was described under the data transformation section because EFC data were sourced from a different database, called Free Application for Federal Student Aid (FAFSA). For that reason, EFC data were fed to Slate and merged with student application data once it is sourced from the FAFSA database. The FAFSA application required students to provide EFC information which can be used to estimate their socioeconomic status. Students who did not submit FAFSA were shown to have no data (i.e., NULL) under this field. In this study, EFC ranged from $0 to $999,999.

Institutional Financial Aid

Institutional scholarships and grants were considered as institutional financial aid. Once students complete and submit their applications, the institution evaluates their applications for admission. Then, the institution determines the total number/amount of scholarships and/or grants and assigned them to students. This was determined based on various metrics, such as studnets' high school academic achievement, superscores, and socioeconomic status. This study focused on the amount of aid and examines how much it impacts students' enrollment decisions. The amount of aid that was offered to students ranges from $0 to $76,762 per academic year.

Proximity

Students were required to provide their mailing/permanent addresses by the time of their application submission. Once their address information were stored in the student database warehouse, Slate automatically calculates the geographic proximity (i.e., distance) from students' addresses to college based on their address zip code. Based on the dataset for the current study, students' proximity ranged from 1 mile to 9,361 miles.

Unemployment Rate

Similar to EFC, unemployment rate data were sourced from Bureau of Labor Statistics (BLS) database. The unemployment rate is calculated as the proportion of unemployed individuals in the labor force, which includes both employed and unemployed individuals. Since the data covers the cohort years from 2013 to 2019, the average unemployment rate trend was also captured from 2013 to 2019. The average unemployment rate range for over seven years was from 3.7% to 7.6%.

Inflation Rate

Inflation rate data was also sourced from BLS database. To monitor the official inflation rate, the consumer price index (CPI) was utilized to measure changes in the cost of living over a period of time. In other words, CPI tracked the average rate of change in U.S. inflation over time. Therefore, the current study reflects CPI trends which ranged from 1.8% to 2.2% over 7 years from 2013 to 2019.

Data Split and k-fold Cross-Validation

Once the data transformation was completed, there were 50 metrics excluding the "enrollment decision" metric. This meant 50 independent features that acted as predictors were inputted into the model to infer the value of “enrollment decision” (the output/dependent variable). The data was split based on the cohort year and the application, admit, and enrolled counts in each cohort year is shown in Table 8.

Table 8. Application Funnel Overview from Fall 2013 to Fall 2019.

Cohort Term	Application N	Admit N	Enroll N
2013	20,554	13,128	2,496
2014	24,148	12,931	2,292
2015	25,478	15,356	2,194
2016	26,806	16,482	2,626
2017	27,528	16,639	2,658
2018	29,176	17,064	2,774
2019	29,807	17,198	2,636

Data Split

The cohort years 2013 to 2018 student records were gathered from the complete dataset to form the training dataset, which was utilized to train, validate and evaluate the prediction performance of the models. The remaining student records from the cohort year 2019 were used to create a test dataset. Out of the combined train-test dataset, comprising 91,600 admitted student records, with 15,040 enrolled and 76,560 non-enrolled students, approximately 84% was used to train and validate the four machine learning algorithm models. The remaining 16%, which included 17,198 admitted student records, with 2,636 enrolled and 14,562 non-enrolled students, was employed to test and assess the prediction performance of the chosen model.

This approach of dividing the entire input data into training and testing datasets is known as the "holdout" method, where the typical split rate is around 80% for training and 20% for testing (Stone, 1977; Nguyen et al., 2021). However, the holdout method's disadvantage is that the test dataset results are greatly influenced by how the researcher classifies the initial data (Yadav

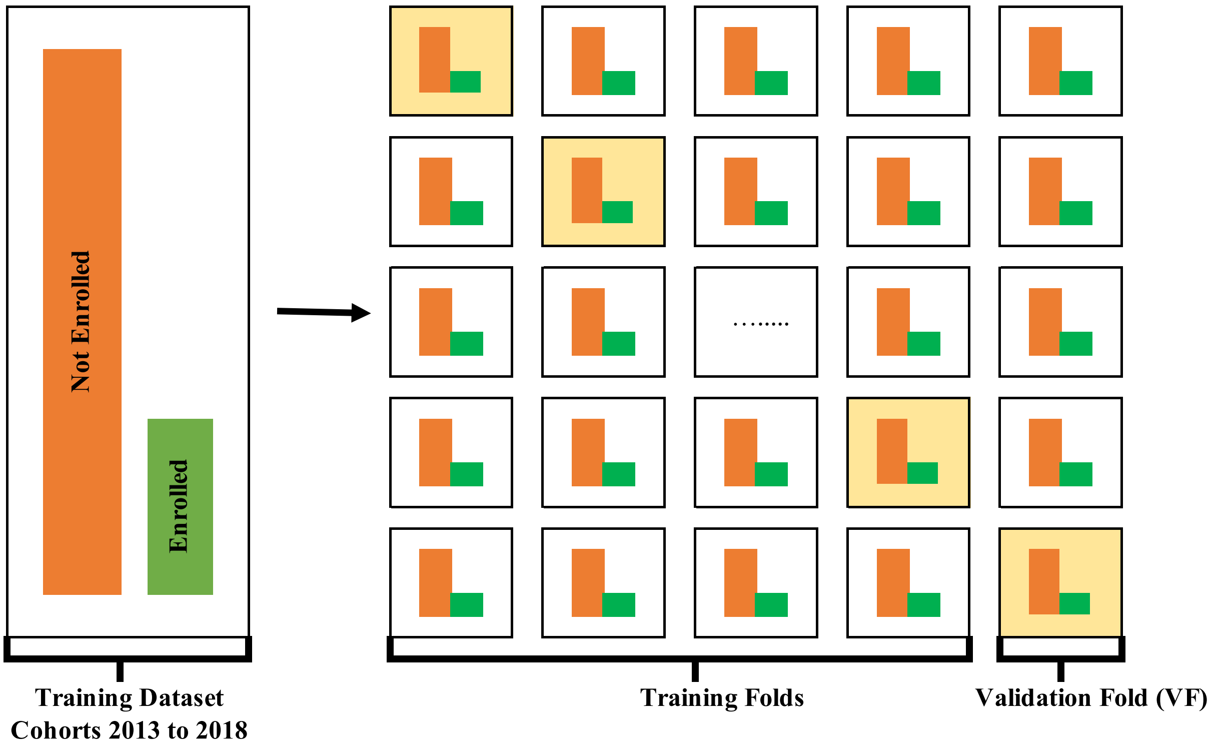
& Shukla, 2016). In other words, a larger percentage of the test dataset may make the model prone to errors as it has less training experience, while a smaller percentage of the test dataset could give the model an unwanted bias towards the training data, leading to underfitting/overfitting of the prediction model. Therefore, data scientists have developed k-fold and stratified k-fold cross-validation methods. For this study, the stratified k-fold cross-validation method was used.

k-fold and Stratified k-fold Cross-Validations

The k-fold cross-validation method involves randomly dividing the training dataset into k subsets, with one subset used as the validation set and the remaining k-1 subsets combined to form a training set (Wong & Yang, 2017; Parker et al., 2007). The machine learning models are then trained k times, with each iteration using a different subset as the validation set and the other subsets as the training set. The error and accuracy estimates are averaged over all k trials to determine the overall performance of the model. This method greatly reduces prediction error by using most of the data for fitting and helps to remove biases by repeating the subsets of the training dataset.

Stratified k-fold cross-validation is a modified version of k-fold cross-validation that is more effective when dealing with datasets that have an imbalanced response variable (Kohavi, 1995; Zeng & Martinez, 2000). It was similar to k-fold cross-validation. But instead of splitting the dataset randomly to k-fold, stratified k-fold spitted the dataset in a way that each fold has the same class distribution at the training dataset (Zeng & Martinez, 2000; Olson & Delen, 2008). For example, the outcome of the dataset shown in Figure 6 is students' enrollment decisions for cohorts from Fall 2013 to Fall 2018. It shows that about five times more students did not enroll than enrolled in the outcome class. When the training dataset was split with a stratified k-fold, each fold would have similar outcome class distribution or five times more not enrolled students than enrolled students.

Figure 6. Stratified k-fold Cross-Validation for Training Data (Cohorts from Fall 2013 to 2018)

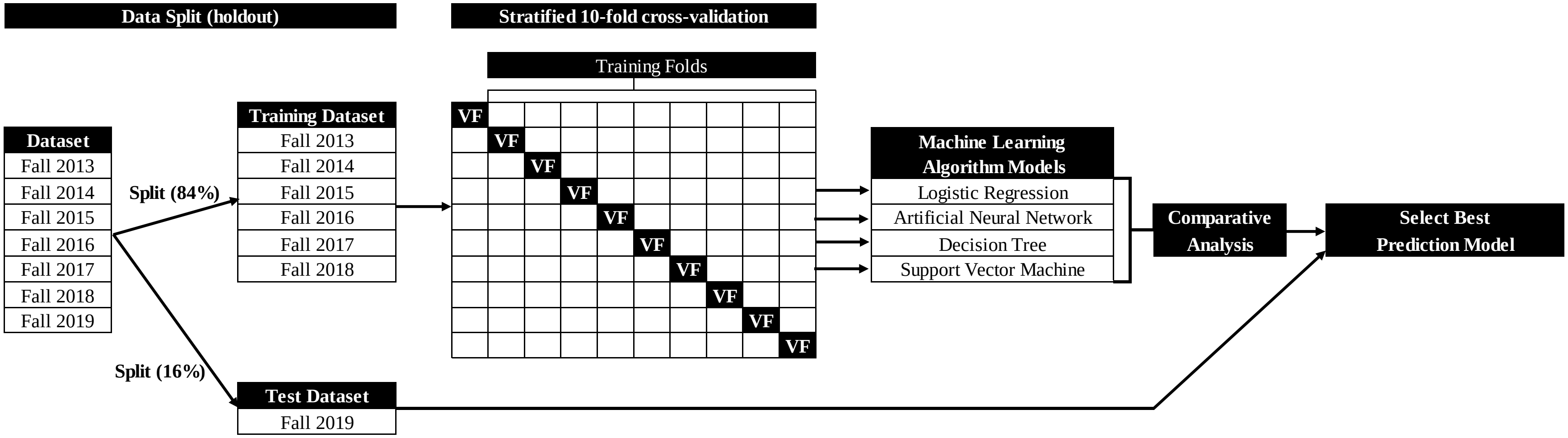


Many researchers suggest that the typical value of k for real-world datasets is 10. In this approach, the data is divided into 10 subsets and during each run, 9 subsets (equivalent to 90% of the data) are used for training while the remaining 1 subset (i.e., 10% of the data) is used for validation (Kohavi, 1995; Kuhn & Johnson, 2013; James et al., 2017).

For illustration purposes, Figure 7 shows how the stratified k-fold cross-validation method was applied to the original data set of the current study. First, the original data was split into two datasets (i.e., training and testing) based on the holdout method. The training dataset involved the cohort data from Fall 2013 to Fall 2018, and the test dataset involved the cohort data from Fall 2019. Then, a stratified 10-fold cross-validation method was implemented in the training dataset which divided the training dataset into 10 subsets. Hence, the 82,440 students including 68,904 non-enrolled and 13,536 enrolled students were used as a training dataset, a combined nine subsets. For the rest of one subset, 9,160 students were used as a validation dataset, including 7,656 non-enrolled and 1,504 enrolled students. This was iterated ten times by capturing the validation dataset ten times independently.

Selected four machine learning models (LR, SVM, DT, and ANN) were trained based on nine training subsets, and the remaining validation subset was rotated k times. After evaluating the four ML models' prediction performances, the best prediction performance model was selected based on comparative analysis. Then, the test dataset (i.e., Fall 2019) was applied to the chosen model and its prediction performance was evaluated and utilized.

Figure 7. The Data Split and 10-fold Cross-Validation Procedure of Train and Test data for Compile Machine Learning



Experiments in Testing Machine Learning Models

Once the train, validation, and test datasets were created from the original data, four machine learning models (i.e., LR, DT, ANN, and SVM) were compiled using train and validation data with stratified 10-fold cross-validation applied. Then, I evaluated each model's classification prediction performance in terms of accuracy, sensitivity, specificity, precision, F_1 score and AUC values. These evaluation metrics were generated from concatenated confusion matrix and mean ROC and PR plots. I chose the best model(s) based on these seven-evaluation metrics and implemented the test data to those model(s). For compiling models, 50 pre-processed independent metrics including one-hot encoding were applied. However, one of the disadvantages of one-hot encoding is that it may lead to a dummy variable trap, leading to a multicollinearity issue. Although DT and ANN are free from multicollinearity concerns (Kotsiantis, 2013; Hansen & Sargent, 2001; Dreiseitl & Ohno-Machado, 2002; Bejou et al., 1996; Tu, 1996), this issue is significant for logistic regression and linear SVM models. Therefore, multicollinearity was considered and applied to both the train/validation and test dataset so that the results can be compared apples to apples. I used a variance inflation factor (VIF) to measure the amount of multicollinearity of 46 independent variable predictors by omitting three one-hot encoded largest variables in each categorical field; Race, Sex, Parent Education Leve (PEL), and High School Academic Climate (HSAC) (Wissmann et al., 2014). In other words, 'White,' 'Female,' 'Parent Education Level with Bachelors Degree,' and 'HSAC_79' were omitted. These three variables became the references/baseline category, and the other dummy variables represent the differences between the remaining and reference categories. Therefore, 46 independent metrics, including dummy variables, were reflected in four machine-learning models.

In addition, I conducted feature importance and prediction probability analysis for chosen model(s). These analyses were conducted to show how the selected model(s) were used pragmatically. Feature importance analyses were useful in identifying the most influential factors toward students' enrollment decisions. Hence, this analysis showed each factor's influential magnitude on the outcome (i.e., student enrollment decisions). Since this is analyzed based on a non-parsimonious setting, it could provide insights into non-influential factors but could impact developing institution's diversity of student body. In addition to feature importance, prediction probability analysis was conducted to measure the probability of students' likelihood to enroll or not at an individual level. Since identifying the best machine learning model was the key of the study, the following subsections described the confusion matrix and ROC plot, along with how those six evaluation metrics were calculated.

Confusion Matrix

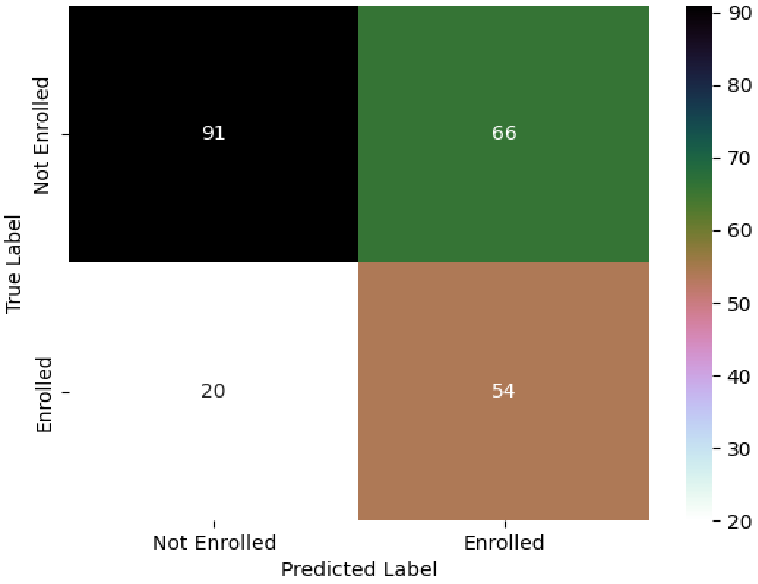
Confusion matrices, which include metrics such as true positives, true negatives, false positives, and false negatives, are commonly used in machine learning algorithms to represent the model's performance on a given dataset (Kohavi & Provost, 1998; Caelen, 2017). These matrices are essential in classification problems as they provide a clear understanding of how often the model correctly predicted a true or false value (Kohl, 2012). Furthermore, the confusion matrix is an effective tool for assessing model performance and can be used to compute other metrics (Kohl, 2012; Sammut & Webb, 2017). As shown in Figure 8, a confusion matrix is constructed using actual and predicted values and is composed of statistics that are calculated to evaluate the model's performance.

Figure 8. Confusion Matrix for Binary Classification of Enrollment Decision

		True Class	
		Enrolled	Not Enrolled
Predicted Class	Enrolled	True Positive (TP)	False Positive (FP)
	Not Enrolled	False Negative (FN)	True Negative (TN)

When attempting to predict dichotomous outcomes such as college enrollment, each observation in a test dataset results in one of four categories: True Positive (TP), False Positive (FP), True Negative (TN), or False Negative (FN) (Guyon & Eliseff, 2007; Kohl, 2012). True positives occur when the model accurately predicts students who have enrolled in college. False positives arise when the model predicts that a student will enroll, but they do not. True negatives arise when the model accurately predicts students who did not enroll in college. False negatives occur when the model predicts that a student will not enroll in college, but they actually enroll. Figure 6 provides an example to illustrate the concepts of the confusion matrix representation. The matrix was generated from a total population of 231 graduate students who were admitted to the Master's Business programs in Fall 2018 at a private university in the Midwest. The predictions made in this example represent whether the students enrolled or not. Figure 9 displays the confusion matrix for a binary classification problem with two classes, namely "enrolled" and "not enrolled".

Figure 9. Example of Confusion Matrix



Many advanced models included confusion matrix features. This feature allowed the model to automatically adjust the model weight and continue calibrating as new data which was collected based on the values derived from the confusion matrix (Sammut & Webb, 2017). This study utilized metrics to report the best-fitting model(s) during the comparison phase.

Classification Accuracy, Sensitivity, Specificity, and Precision

When working within supervised learning states, the four main components of measuring a model fit were model accuracy, sensitivity, specificity, and precision (Brieman, 1984). These four components were calculated using the confusion matrix values (Kohl, 2012).

Classification accuracy is defined as the proportion of the number of correct predictions to all the predictions made by the model (Geron, 2017). To evaluate the model's accuracy in distinguishing between students who enrolled and those who did not, the ratio of the sum of true-positive (TP) and true-negative (TN) predictions to the sum of all evaluated observations (i.e., the total size of the predicted population) was calculated.

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

Sensitivity refers to the probability of a prediction being true when the actual class is true. Simply, it describes how well the model can predict positive instances. It is also referred to as "True positive rate" or "Recall" and is calculated as the ratio of true positives to the actual positive cases. To estimate the model's sensitivity in predicting college enrollment (while ignoring correct predictions of students not enroled in college), the ratio of true positives (TP) to the sum of true positives or false negatives (FN) of student enrollment was calculated.

$$Sensitivity\ (i.e., Recall) = \frac{TP}{TP+FN}$$

Specificity refers to the probability of the model's prediction being false when the actual class is false. In other words, it describes how specific the model is when predicting negative instances. Specificity is calculated as the ratio of true negatives to actual negative cases. To measure the model's specificity in predicting students who did not enroll in college (ignoring successful predictions of students enrolling in college), the ratio of true-negative (TN) students correctly predicted as not enrolling in college was calculated from the sum of true-negative (TN) and false-positive (FP) students who did not enroll in college.

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

In order to have a comprehensive evaluation of the model's performance, it is not sufficient to focus only on the true positive rate (i.e., sensitivity) and true negative rate (i.e., specificity). The concept of precision, which measures the positive predicted value, is also important to consider. Precision describes the probability of the prediction being correct when the model identifies it as positive, and it reflects how precise the model is when predicting positive instances. To calculate precision, we divide the number of true positives (TP) by the total number of positive predictions, which is the sum of true positives (TP) and false positives (FP).

$$Precision = \frac{TP}{TP+FP}$$

F_1 Score

When using classification models in machine learning, one of the common metrics that was also used to assess the quality of the model is F_1 score (Chicco & Jurman, 2020). A model's F_1 score was calculated by taking a harmonic mean of its precision and sensitivity. It finds the most optimal balanced confidence score threshold where precision and recall give the highest F_1 score. If the F_1 score is high, precision and recall are high, and vice versa. An equation for calculating F_1 score is shown below. The range of F_1 Score is from 0 to 1 and the closer it is to 1, it is determined to be the better model.

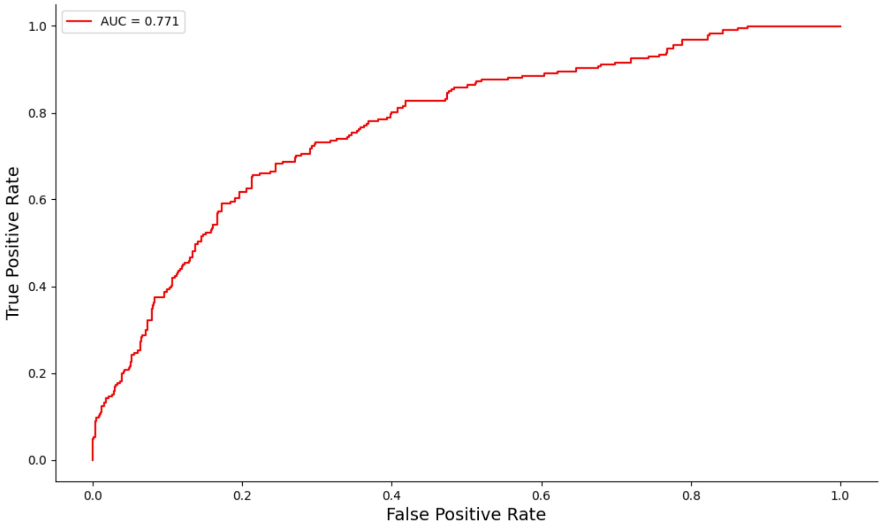
$$F_1 = \frac{2 * (Precision * Sensitivity)}{Precision + Sensitivity}$$

Receiver Operating Characteristics (ROC) and Area Under Curve (AUC)

When dealing with binary classification problems, the typical practice was to utilize a probability threshold of 0.5 for classification predictions (Arisholm et al., 2010). However, in some situations, an alternate threshold may be more appropriate. The Receiver Operating Characteristics (ROC) curve is the most widely used technique for displaying the performance of a binary classifier at various thresholds (Fawcett, 2006; Muschelli, 2020). It is created by plotting the True Positive Rate versus the False Positive Rate, with the latter calculated as 1-Specificity. The Area Under the Curve (AUC) was determined from the ROC curve, indicating the likelihood that a classifier model would rank a randomly chosen positive instance higher than a randomly chosen negative one (Fawcett, 2006).

Figure 10 displays the ROC curve generated from the example data obtained from the confusion matrix illustrated in Figure 9. The area under the red line in the plot represents the AUC value of the classifier, which, in this case, was 0.759. AUC values above 0.9 are considered excellent, between 0.8 and 0.9 as good, 0.7 and 0.8 as fair, 0.6 and 0.7 as poor, and 0.5 and 0.6 as failed (Ludemann et al., 2006; Obuchowski, 2003; Khouli et al., 2009). Generally, the ROC curve of a model closer to the upper-left corner (i.e., a value closer to 1) is deemed to exhibit better classification performance (Fawcett, 2006; Japkowicz, 2013).

Figure 10. Example of Receiver Operating Characteristics Curve Plot

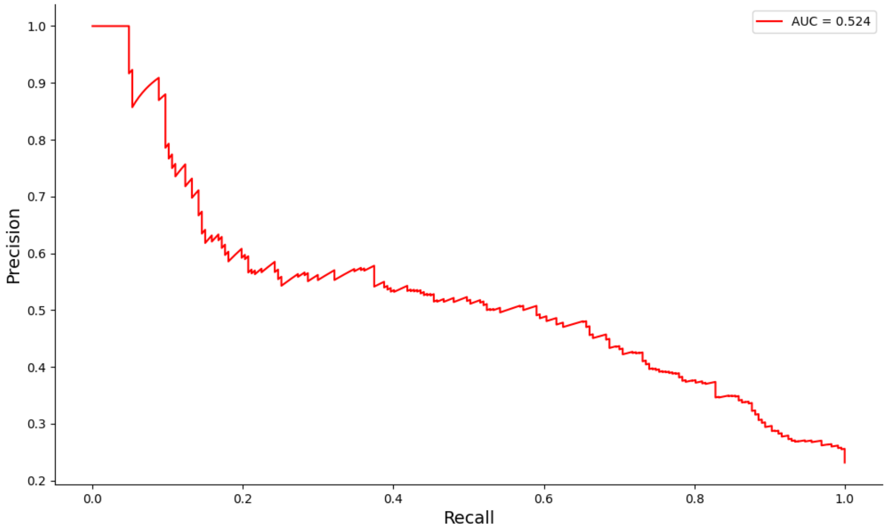


Precision-Recall (PR) Curve and Area Under the Curve (AUC)

Although the Receiver Operating Characteristics (ROC) curve is a widely used and effective method to evaluate binary classification models (Fawcett, 2006; Muschelli, 2020), it can be unreliable when dealing with heavily imbalanced data. Davis & Goadrich (2006) have noted that the ROC curve tends to overestimate the classification performance of a model in such cases. To address this limitation, the Precision-Recall (PR) curve has been identified as a useful alternative for evaluating model performance (Fu et al., 2018). The PR curve depicts the relationship between precision and recall, with precision values (TP/(TP+FN)) plotted on the y-axis and recall values (TP/(TP+FP)) on the x-axis. As precision is also known as positive predicted value and recall as true positive rate, the PR curve is particularly useful in classifying true positive cases rather than negative cases (Bekkar et al., 2013; Japkowicz, 2013).

Figure 11 presents an example PR curve constructed from the confusion matrix data shown in Figure 9. The area under the red line in the plot represents the AUC (area under the curve) of the classifier, which in this case was 0.524. As previously mentioned, an AUC value between 0.9 and 1 is considered excellent, between 0.8 and 0.9 as good, 0.7 and 0.8 as fair, 0.6 and 0.7 as poor, and 0.5 and 0.6 as failed (Ludemann et al., 2006; Obuchowski, 2003; Khouli et al., 2009). Generally, a PR curve that approaches the top-right corner (i.e., a value closer to 1) is indicative of better classification performance (Cavert & Khoshgoftaar, 2019; Japkowicz, 2013).

Figure 11. Example of Precision-Recall Curve Plot



Feature Importance Analysis

After selecting the best ML algorithm model, drop-column importance was performed to determine the impact of each independent variable on the model's classification performance. This method is straightforward as it assesses the significance of independent variables/features by comparing a model that includes all features with one that excludes certain features (Chen et al., 2020). This approach is highly accurate in measuring feature importance (Saarela & Jauhiainen, 2021), but it requires a significant amount of computational time since the model needs to be retrained for each variant of the dataset (by removing one feature column at a time) (Chen et al., 2020; Saarela & Jauhiainen, 2021).

Potential Predictve Probability Analysis

Apart from conducting feature importance analysis, potential predictive probability analysis was also carried out to demonstrate the practical utilization of the selected model(s) in supporting the institution's enrollment management plan and strategies. This analysis involved compiling the enrollment probability of individual students using the selected models, which were determined through comparative analysis. This process enables the institution to gain insights into the enrollment likelihood of each admitted student and develop personalized strategies for them efficiently.

Summary

This chapter covered various aspects of the data collection process, data pre-processing (including one-hot encoding and stratified k-fold cross-validation), the issue of the dummy variable trap leading to multicollinearity, and a narrative summary of the methodology for comparative analysis (including confusion matrix, accuracy, sensitivity, specificity, precision, F_1 score, ROC-AUC, and PR-AUC). The chapter then explored four supervised machine learning models (Artificial Neural Network, Decision Tree, Support Vector Machine, and Logistic Regression) to develop a model that accurately predicts students' enrollment decisions. Additionally, feature importance and potential predictive probability analyses are discussed, showcasing the practical utilization of the selected model(s) to support enrollment management at the institution.

The next chapter presents the results of each machine learning model and identifies the optimal models for predicting students' enrollment decisions. It also demonstrates how the selected models were practically implemented.

CHAPTER FOUR

RESULTS

This chapter discusses the results beginning with a description of the dataset and then analyzes the findings related to the three research questions mentioned in Chapters 1 and 3. This study applied four selected machine learning algorithms (i.e., LR, SVM, DT, and ANN) to a preprocessed dataset containing 108,798 admits, including 17,676 enrolled and 91,122 non-enrolled students and prediction models were compiled. As mentioned in the methodology chapter, 46 independent metrics were applied to the models, taking multicollinearity into account. These prediction models were created using the “scikit-learn” package in Python 3.10 and were fitted to the training data. Using stratified 10-fold cross-validation, trained models were used to predict the validation data. Confusion matrices related to the four algorithm models were generated from the predictions of the validation data.

Descriptive Statistics

The original dataset was split into two: train/validation and test datasets based on the cohort terms along with stratified 10-fold cross-validation applied. The train/validation dataset included the admit data (i.e., enrolled and non-enrolled students) from Fall 2013 to Fall 2018 whereas test data included the ones for Fall 2019. Hence, a total of 108,798 admits, including 17,676 enrolled (16%) and 91,122 non-enrolled (84%) students were in the train/validation dataset. For the test dataset, a total of 17,198 admits, including 2,636 enrolled (15%) and 14,562 non-enrolled (85%) students were in the test dataset. Students’ academic profiles and a break down by their demographic, socioeconomic, geographic, and high school academic climate attributes are summarized in Tables 9 and 10. Overall, approximately three-fifths and two-thirds of enrolled students were white and female, respectively. At the parent educational level, more than 70% of the enrolled students' parents had bachelor's and/or graduate degrees. For the high school academic climate cluster (HSAC), the majority of enrolled students were from three of the twenty-nine unique clusters:

During seven cohort years from 2013 to 2019, 29% to 37% of enrolled students came from high schools coded as cluster 79 which reflected that those high schools were the 'public schools primarily serving highly educated, middle-class populations. Another 19% to 24% of enrolled students were from high schools coded as cluster 70 which those high schools were the 'public schools primarily serving affluent suburban populations.' Lastly, the other 19% to 26% of enrolled students were from high schools coded as cluster 68 which those high schools were the 'religious schools primarily serving catholic populations.' These trends were consistent for the non-enrolled student pool across terms. It showed that 25% to 37% of non-enrolled students were coming from high schools coded as cluster 79, 15% to 23% of non-enrolled students were coming from cluster 70, and 18% to 26% of non-enrolled students were coming from cluster 68. In addition, Table 9 shows that the averages of expected family contribution and the amount of institutional financial aid were higher for enrolled students than non-enrolled students across terms.

Excluding the Fall 2013, 2016 and 2017 terms, the average standardized test scores for the enrolled student pool were slightly lower than the non-enrolled student pool (Fall 2014: 26.88 < 27.27; Fall 2015: 26.52 < 26.74; Fall 2018: 26.70 < 26.92; Fall 2019: 26.91 < 27.19).

Also, excluding Fall 2013, the average high school GPAs for the enrolled student pool were slightly lower than the non-enrolled student pool (Fall 2014: 3.82 < 3.85; Fall 2015: 3.76 < 3.82; Fall 2016: 3.78 < 3.84; Fall 2017: 3.80 < 3.83; Fall 2018: 3.80 < 3.87; Fall 2019: 3.84 < 3.92). In Fall 2016 and 2017, the average proximity (i.e., distance from college to student's residence) was slightly higher for the enrolled student pool compared to the non-enrolled student pool (Fall 2016: 295.0 > 252.0; Fall 2017: 364.3 > 360.4).

Table 9. Enrolled vs. Not Enrolled Students’ Demographic Overview

	Train/Validation Dataset												Test Dataset	
	Fall 2013		Fall 2014		Fall 2015		Fall 2016		Fall 2017		Fall 2018		Fall 2019	
	Enrolled	Not Enrolled	Enrolled	Not Enrolled	Enrolled	Not Enrolled	Enrolled	Not Enrolled	Enrolled	Not Enrolled	Enrolled	Not Enrolled	Enrolled	Not Enrolled
	(n=2,469)	(n=10,632)	(n=2,292)	(n=10,639)	(n=2,194)	(n=13,162)	(n=2,626)	(n=13,856)	(n=2,658)	(n=13,981)	(n=2,774)	(n=14,290)	(n=2,636)	(n=14,562)
Race														
White	59.9%	58.2%	58.8%	60.0%	56.8%	54.9%	60.0%	54.7%	57.6%	54.5%	56.9%	53.2%	58.8%	53.4%
Black or African American	3.1%	5.0%	4.7%	3.1%	4.7%	5.5%	5.0%	6.3%	3.9%	6.2%	4.7%	5.7%	4.7%	5.4%
Hispanic or any race	15.2%	18.1%	17.6%	14.9%	15.6%	20.3%	15.0%	21.0%	17.3%	20.9%	17.6%	22.6%	18.1%	22.4%
Asian	15.9%	11.9%	11.9%	14.4%	16.0%	13.3%	14.0%	12.4%	15.2%	12.7%	15.0%	12.6%	12.4%	13.3%
Other	5.0%	5.1%	5.6%	6.1%	5.2%	4.3%	4.8%	4.3%	5.0%	4.7%	5.0%	5.0%	5.5%	4.7%
Unknown	1.0%	1.7%	1.4%	1.6%	1.5%	1.6%	1.2%	1.3%	1.0%	1.1%	0.8%	0.9%	0.6%	0.8%
Sex														
Female	66.0%	67.6%	67.9%	68.3%	67.0%	69.7%	68.7%	70.6%	68.4%	68.5%	68.0%	70.2%	67.0%	69.5%
Male	34.0%	32.4%	32.1%	31.7%	33.0%	30.3%	31.3%	29.4%	31.6%	31.5%	32.0%	29.8%	33.0%	30.5%
Parent Education Level														
Some High School	3.5%	3.5%	3.2%	3.2%	3.3%	3.6%	3.0%	5.2%	4.0%	6.2%	3.2%	4.9%	2.9%	5.0%
High School Graduate	9.9%	7.9%	7.8%	8.7%	7.8%	9.4%	9.6%	17.8%	7.7%	10.1%	9.1%	9.8%	8.2%	8.9%
Associate Degree	5.7%	4.3%	4.8%	4.2%	5.0%	4.1%	4.9%	7.3%	3.5%	4.4%	4.7%	4.4%	4.4%	4.1%
Some College	10.2%	10.5%	10.9%	9.7%	12.2%	10.5%	11.4%	9.9%	9.9%	9.0%	9.6%	8.9%	8.2%	8.2%
Bachelors Degree	37.2%	38.2%	39.8%	38.2%	36.7%	36.9%	36.2%	31.4%	36.4%	36.5%	37.2%	35.6%	36.0%	36.0%
Graduate Degree	33.4%	35.6%	33.5%	36.1%	35.0%	35.4%	34.9%	28.3%	38.5%	33.8%	36.2%	36.3%	40.3%	37.8%
Standardized Test	26.90	26.71	26.88	27.27	26.52	26.74	26.73	26.25	26.85	26.73	26.70	26.92	26.91	27.19
Superscore Mean (SD)	(3.35)	(3.62)	(3.35)	(3.70)	(3.64)	(3.78)	(3.65)	(4.13)	(3.54)	(3.83)	(3.72)	(4.03)	(3.94)	(4.11)
High School	3.78	3.76	3.82	3.85	3.76	3.82	3.78	3.84	3.80	3.83	3.80	3.87	3.84	3.92
GPA Mean (SD)	(0.47)	(0.47)	(0.46)	(0.45)	(0.48)	(0.48)	(0.47)	(0.51)	(0.48)	(0.48)	(0.48)	(0.47)	(0.49)	(0.47)
Proximity Mean	326.8	517.3	334.0	363.8	330.4	365.9	295.0	252.0	364.3	360.4	309.0	379.9	355.0	422.7
Instituiona Financial Aid Mean	$15,071.40	$14,230.70	$15,599.63	$15,392.24	$16,315.65	$16,107.73	$19,150.94	$17,913.47	$19,670.24	$19,497.61	$20,400.44	$20,179.55	$20,831.59	$20,521.02
Expected Family Contribution Mean	$22,185.72	$17,092.32	$34,182.69	$25,514.77	$36,647.57	$22,403.21	$33,450.15	$21,856.54	$35,385.96	$27,253.53	$39,336.27	$30,803.94	$46,143.25	$38,053.25
Unemployment Rate	7.6%		6.5%		5.4%		4.9%		4.5%		4.0%		3.7%	
Inflation Rate	1.8%		1.8%		1.8%		2.2%		1.9%		2.0%		2.2%	

Table 10. Enrolled vs. Not enrolled by High School Academic Climate Overview

High School Academic Climate	Train / Validation Dataset												Test Dataset	
	Fall 2013		Fall 2014		Fall 2015		Fall 2016		Fall 2017		Fall 2018		Fall 2019	
	Enrolled (n=2,469)	Not Enrolled (n=10,632)	Enrolled (n=2,292)	Not Enrolled (n=10,639)	Enrolled (n=2,194)	Not Enrolled (n=13,162)	Enrolled (n=2,626)	Not Enrolled (n=13,856)	Enrolled (n=2,658)	Not Enrolled (n=13,981)	Enrolled (n=2,774)	Not Enrolled (n=14,290)	Enrolled (n=2,636)	Not Enrolled (n=14,562)
51	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
52	0.2%	0.4%	0.1%	0.6%	0.1%	0.5%	0.2%	0.5%	0.4%	0.3%	0.4%	0.4%	0.2%	0.4%
53	0.2%	0.3%	0.2%	0.1%	0.2%	0.1%	0.0%	0.0%	0.2%	0.2%	0.2%	0.2%	0.1%	0.2%
54	0.1%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%
55	0.0%	0.3%	0.1%	0.2%	0.2%	0.1%	0.2%	0.0%	0.2%	0.1%	0.1%	0.2%	0.3%	0.3%
56	0.0%	0.1%	`	0.1%	0.2%	0.1%	0.0%	0.0%	0.1%	0.1%	0.1%	0.1%	0.2%	0.0%
57	0.4%	1.0%	0.8%	1.3%	0.9%	1.0%	0.8%	1.6%	1.1%	2.1%	0.8%	1.9%	1.0%	1.5%
58	0.1%	0.0%	0.1%	0.2%	0.2%	0.1%	0.1%	0.0%	0.2%	0.2%	0.1%	0.1%	0.3%	0.1%
59	1.0%	1.1%	0.9%	1.3%	1.0%	0.9%	0.6%	0.5%	1.2%	1.2%	0.7%	1.1%	0.9%	0.9%
60	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.1%	0.0%	0.2%	0.1%	0.1%	0.0%
61	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
62	0.9%	1.1%	0.7%	0.9%	0.9%	1.2%	0.4%	1.6%	0.5%	0.8%	0.6%	0.8%	1.1%	1.1%
63	0.1%	0.3%	0.1%	0.1%	0.1%	0.1%	0.0%	0.0%	0.2%	0.3%	0.1%	0.3%	0.1%	0.2%
64	0.4%	0.7%	0.4%	0.4%	0.4%	0.5%	0.4%	0.0%	0.3%	0.5%	0.5%	0.6%	0.3%	0.5%
65	6.4%	7.9%	7.0%	7.4%	6.3%	7.0%	6.9%	8.9%	7.3%	7.9%	7.6%	8.5%	7.7%	8.3%
66	0.5%	0.0%	0.3%	0.5%	0.5%	0.6%	0.5%	0.0%	0.5%	0.8%	0.5%	0.9%	0.5%	0.5%
67	6.9%	7.7%	6.9%	7.2%	8.4%	7.9%	8.4%	8.4%	6.7%	8.8%	7.6%	8.8%	6.3%	6.9%
68	22.0%	26.4%	21.7%	22.4%	24.0%	22.6%	19.5%	17.8%	19.7%	20.6%	20.4%	18.6%	19.2%	20.2%
69	0.1%	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%	0.5%	0.1%	0.1%	0.0%	0.1%	0.0%	0.1%
70	18.9%	18.0%	20.3%	20.6%	19.7%	22.8%	24.4%	14.7%	24.6%	21.0%	24.5%	23.3%	25.9%	23.2%
71	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
72	0.6%	0.8%	0.6%	0.4%	0.6%	0.7%	0.6%	0.0%	0.7%	0.6%	1.2%	0.5%	1.2%	0.7%
73	1.1%	2.0%	1.6%	2.5%	1.6%	3.4%	1.8%	4.7%	2.3%	4.0%	1.0%	3.3%	1.1%	3.3%
74	0.2%	0.1%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.1%	0.0%	0.3%	0.2%
75	2.4%	2.5%	2.6%	3.4%	4.0%	3.8%	2.2%	3.7%	4.5%	4.3%	4.1%	4.3%	3.4%	3.6%
76	0.7%	1.3%	0.2%	0.3%	0.3%	0.2%	0.2%	0.0%	0.2%	0.3%	0.5%	0.4%	0.4%	0.3%
77	0.1%	0.1%	0.2%	0.0%	0.0%	0.0%	0.0%	0.5%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
78	0.2%	0.7%	0.3%	0.2%	0.1%	0.2%	0.2%	0.0%	0.0%	0.3%	0.3%	0.2%	0.2%	0.1%
79	36.7%	27.1%	35.0%	29.7%	30.2%	26.0%	32.0%	36.6%	28.8%	25.5%	28.8%	25.2%	29.8%	27.2%

Comparative Analysis

Once the data were ready after applying the pre-processing steps, I wrote python code using the sklearn package to compile four machine learning models (see Appendix C) – Logistic Regression, Decision Trees, Support Vector Machine, and Artificial Neural Network, and implemented train and validation data along with stratified 10-fold cross-validation. As mentioned in the methodology chapter, 46 independent variables were implemented in four models. Three dummy variables were excluded from each three categories representing Race, Sex, Parent Education Leve (PEL), and High School Academic Climate (HSAC) to avoid the dummy variable trap which leads to multicollinearity issue. In short, one-hot coded variables for ‘Race_White’, ‘Sex_Female’, ‘PEL_Bachelors Degree’, and ‘HSAC_79’, with the largest number of data were omitted and used as the references/baseline category for other one-hot encoded dummy variables (Wissmann et al., 2014). To test and confirm for multicollinearity among independent variable predictors, I used a variance inflation factor (VIF) to measure the amount of multicollinearity of 46 independent variable predictors. Table 11 shows that 46 variables did not introduce multicollinearity concerns, with showing VIF values less than 2.5 (Senaviratna & Cooray, 2019). Once the models were compiled, comparative analyses were conducted based on seven calculated metrics; accuracy, sensitivity, specificity, precision, F_1 score, and AUC. The outcomes of each model regarding training and validation data are presented in the next subsections.

Table 11. Multicollinearity Assessment using Variance Inflation Factor (VIF)

Predictors	VIF
SAT_ACT_SuperScore	1.583146
Race_Black_AA	1.240530
Race_Hispanic	1.665556
Race_Asian	1.256161
Race_Other	1.170193
Race_Unknown	1.016480
Sex_Male	1.578803
HSGPA	1.579285
PEL_Some_High_School	1.297855
PEL_High_School_Graduate	1.375282
PEL_Associates_Degree	1.140821
PEL_Some_College	1.317706
PEL_Graduate_Degree	2.023563
HSCluster_51	1.000292
HSCluster_52	1.083763
HSCluster_53	1.008446
HSCluster_54	1.002414
HSCluster_55	1.010819
HSCluster_56	1.003763
HSCluster_57	1.146465
HSCluster_58	1.006437
HSCluster_59	1.042653
HSCluster_60	1.002164
HSCluster_61	1.000880
HSCluster_62	1.090798
HSCluster_63	1.014389
HSCluster_64	1.124949
HSCluster_65	1.332619
HSCluster_66	1.025642
HSCluster_67	1.340699
HSCluster_68	1.933881
HSCluster_69	1.004773
HSCluster_70	1.90962
HSCluster_71	1.000021
HSCluster_72	1.024889
HSCluster_73	1.257415
HSCluster_74	1.004035
HSCluster_75	1.155458
HSCluster_76	1.054432
HSCluster_77	1.003094
HSCluster_78	1.020286
Distance_from_Campus	1.322892
Unemployment_Rate	2.140668
Inflation_Rate	2.018814
Fed_Efc	1.063255
Inst_Fin_Aid	1.665885

Logistic Regression (LR)

I used LogisticRegression() function to compile the LR model using training and validation data. Since a stratified 10-fold cross-validation method was applied, concatenated confusion matrix was generated. To be specific, ten different confusion matrices were generated based on the ten independent validation datasets. Each validation dataset was paired with the rest of the nine pieces of the training dataset and applied to the model. These ten confusion matrices were concatenated to evaluate the classification performance of the training/validation data as a whole. Hence, the values of five indices; accuracy, sensitivity, specificity, precision, and F_1 score were calculated based on the concatenated confusion matrix. The concatenated confusion matrix of the LR model based on training and validated data is shown in Figure 12.

Figure 12. Concatenated Confusion Matrix of LR Model on Train/Validation Data

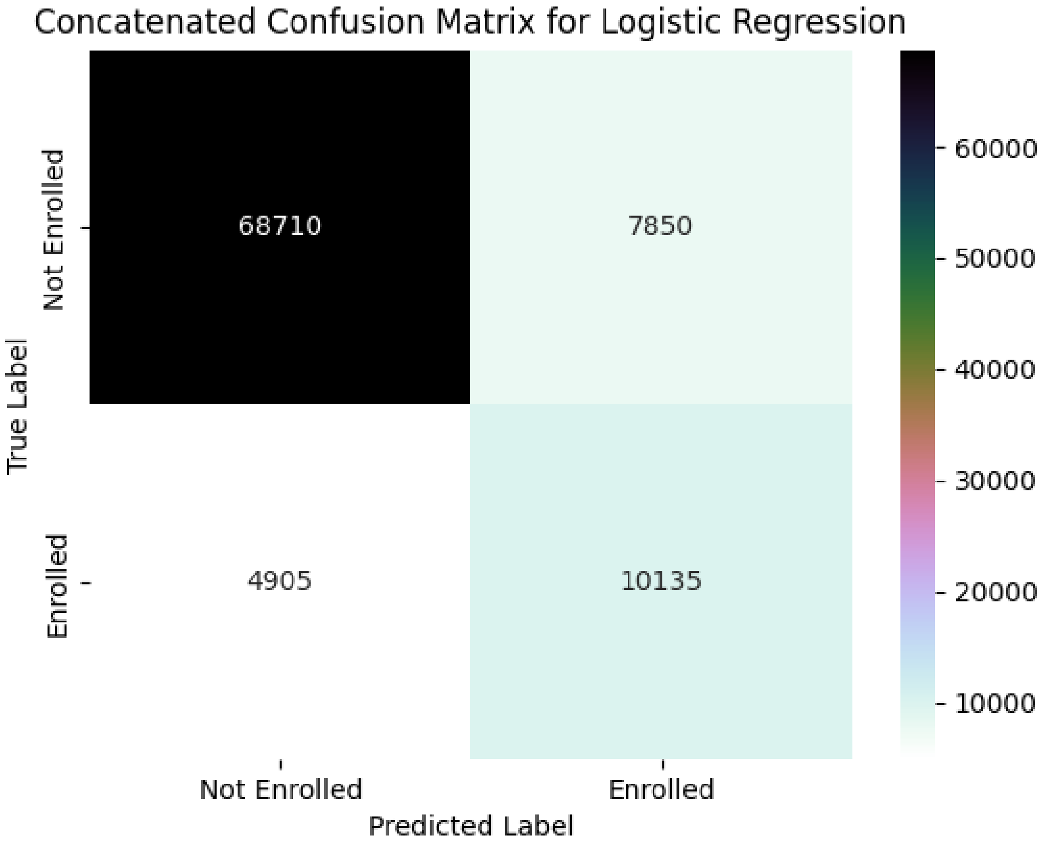


Figure 12 shows that the predicted number of enrolled students who actually enrolled (i.e., true positives) was 10,135, whereas the predicted number of enrolled students who actually did not enrolled was 7,850 (i.e., false positives). Also, the predicted number of non-enrolled students who actually not enrolled was 68,710 (i.e., true negatives) and who actually enrolled was 4,905 (i.e., false negatives). The color bar on the right displays a range of colors that correspond to different values in the confusion matrix. Hence, darker colors represent higher values, while lighter colors represent lower values. Based these outcomes, five evaluation indices of accuracy, sensitivity, specificity, precision, and F_1 score were calculated below.

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} = \frac{68710 + 10135}{68710 + 7850 + 4905 + 10135} = 0.860753 \approx 0.861$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{10135}{10135 + 4905} = 0.673870 \approx 0.674$$

$$Specificity = \frac{TN}{TN + FP} = \frac{68710}{68710 + 7850} = 0.897466 \approx 0.897$$

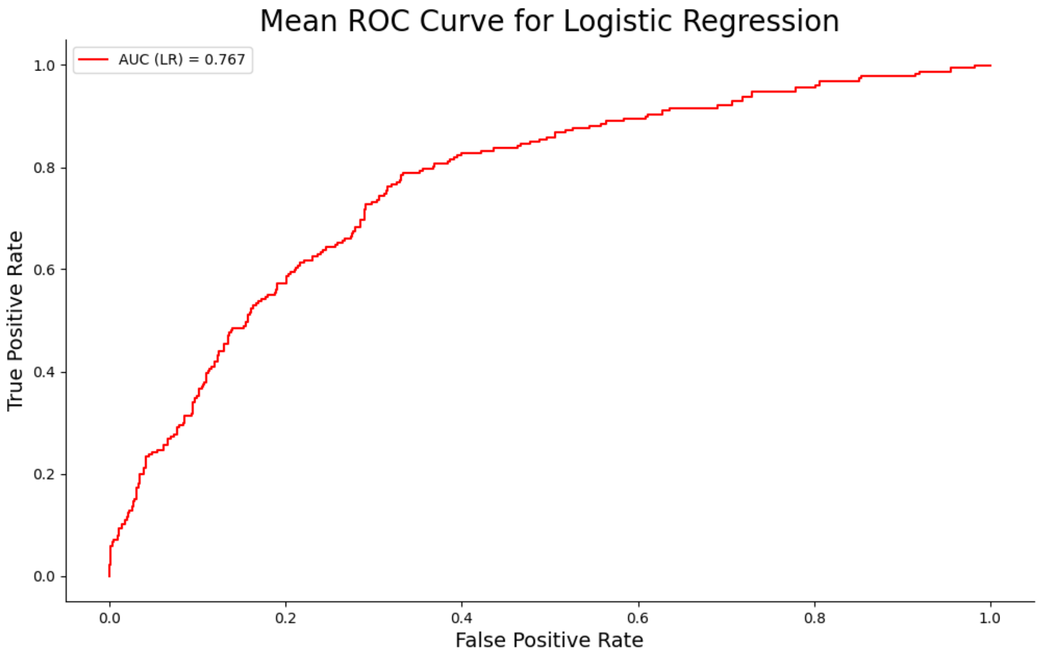
$$Precision = \frac{TP}{TP + FP} = \frac{10135}{10135 + 7850} = 0.563525 \approx 0.564$$

$$F_1 = \frac{2*(Precision*Sensitivity)}{Precision+Sensitivity} = \frac{2*(0.563525*0.673870)}{0.563525+0.673870} = 0.613777 \approx 0.614$$

The overall calculations show that the classification accuracy, sensitivity, specificity, precision, F_1 score of the LR model were 0.861, 0.674, 0.897, 0.564, and 0.614, respectively. The accuracy value of 0.861 implies that the accuracy of the classification performance of LR model is 86.1%. In other words, the LR model's classification performance accuracy for predicting students' enrollment decisions (i.e., enroll and non-enroll) is 86.1%. The sensitivity value of 0.674 implies that the LR model's probability of predicting the students as enrolled when they actually enrolled is 67.4%. Similar to sensitivity, a specificity value of 0.897 shows that the model to predict students as non-enrolled when they actually did not enroll is 89.7%. The precision value of 0.564 implies the ratio of the actual number of enrolled students over the predicted number of enrolled students in the model. This means that 56.4% of students were actually enrolled in the predicted enrolled pool which the model predicted and classified students as enrolled. As mentioned earlier, F_1 score shows the harmonic mean between precision and recall and it is an effective evaluation index for the model's performance on imbalanced data. It ranges between 0 and 1. Although there is no specific threshold for F_1 score to determine whether its classification performance is bad, fair, or good, the model is evaluated to show better performance when F_1 score is closer to 1. The LR model's F_1 score value is 0.614. This is relatively lower than the accuracy (0.861) and specificity (0.897), which gives a hint that the LR model is less effective in showing predictive classification performance for imbalanced data. However, there is no specific threshold of F_1 score to determine the model’s classification performance efficiency. Hence, PR-AUC was generated as a salient index to evaluate the model’s classification performance toward imbalanced data. This is presented after showing the ROC-AUC value and its evaluation.

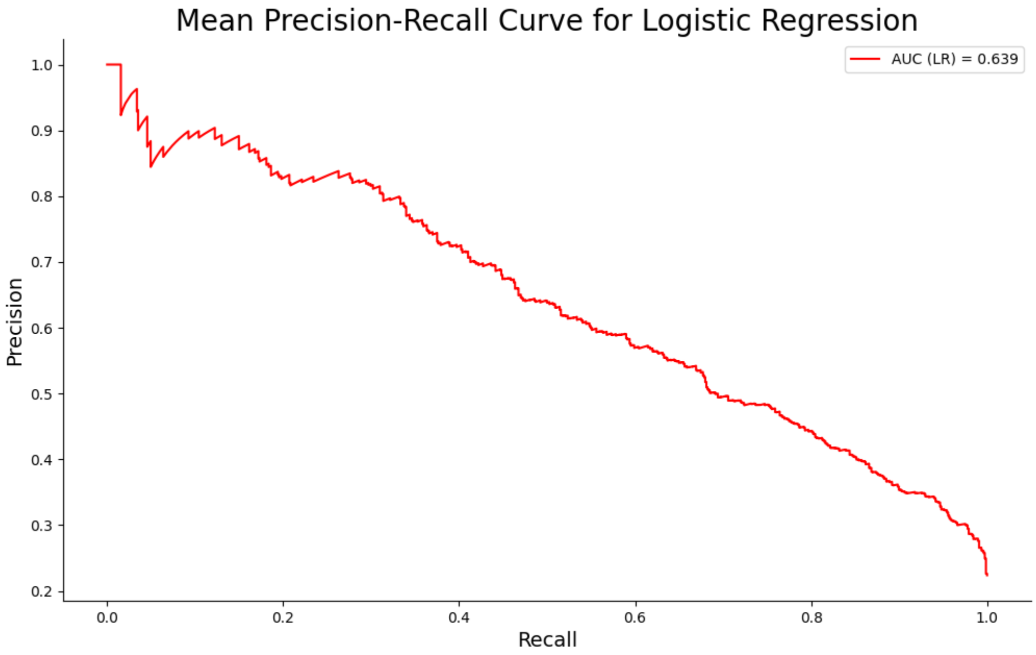
After compiling five comparable metrics based on a concatenated confusion matrix, a mean ROC curve plot was created to determine the Area Under the Curve (AUC) value. As mentioned in the methodology chapter, AUC of ROC curve showed how well the model performs in compiling classification predictions. Since stratified 10-fold cross-validation was applied, the mean of ten ROC curve plots were created and shown in Figure 13. It shows that the average value of AUC is 0.767. As mentioned earlier in the methodology chapter, the model with AUC value between 0.7 and 0.8 were considered to have a fair classification performance. However, this gives an optimistic evaluation with disregarding the data imbalance.

Figure 13. Mean ROC Curve for LR Model on Train/Validation Data with Stratified 10-fold Cross-Validation



Although the ROC curve plot and its AUC provided an insight that the LR model's classification performance is good, Precision-Recall (PR) curve plot and related AUC are also compiled for more rigorous evaluation toward classification performance for imbalanced data, especially focused on classifying true positive cases. As mentioned earlier, PR-AUC is an optimal metric to evaluate the model's classification performance when dealing with highly imbalanced data, particularly when the true positive cases are extremely small. Similar to how the ROC curve was generated, the mean PR curve was generated and shown in Figure 14. It shows that the value of AUC for the mean PR curve of the LR model is 0.639. Since AUC falls between 0.6 and 0.7, the PR model shows poor classification performance on imbalanced data, specifically classifying the true positive cases. Similar to how F_1 score was evaluated, PR-AUC gave a more specific hint that the LR model is not efficient for making classification predictions on students as enrolled who were actually enrolled.

Figure 14. Mean PR Curve for LR Model on Train/Validation Data with Stratified 10-fold Cross-Validation



Decision Tree (DT)

For compiling the DT model, I used the DecisionTreeClassifier() function which includes default conditions such as following the gini impurity criterion with no limitation on the numbers of depths, splits, and leaf nodes. Like generating the LR model, the DT model was compiled using the train and validation data with stratified 10-cross validation applied. Figure 15 shows the concatenated confusion matrix of the DT model based on training and validated data.

Figure 15. Concatenated Confusion Matrix of DT Model on Train/Validation Data

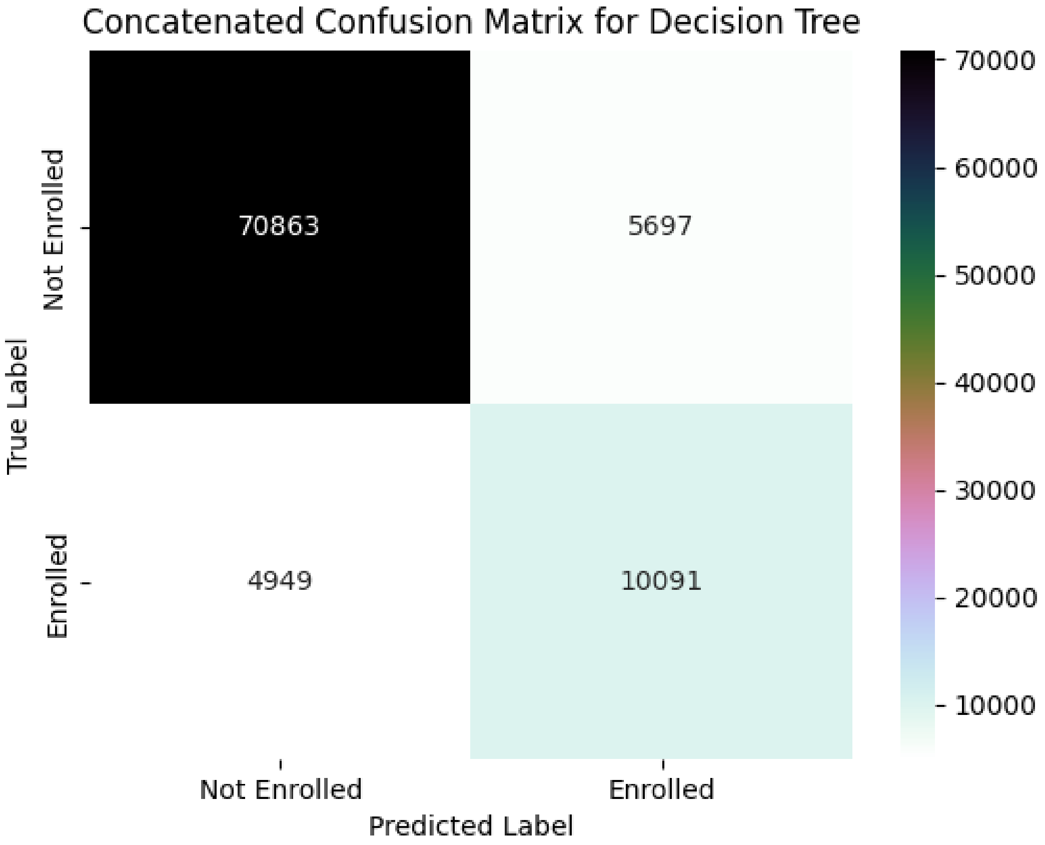


Figure 15 shows that the predicted number of enrolled students who actually enrolled (i.e., true positives) was 10,091, whereas the predicted number of enrolled students who actually did not enroll was 5,697 (i.e., false positives). Also, the predicted number of non-enrolled students who actually did not enroll was 70,863 (i.e., true negatives) and who actually enrolled was 4,949 (i.e., false negatives). Based on these outcomes, five evaluation indices of accuracy, sensitivity, specificity, precision, and F_1 score were calculated as below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{70863 + 10091}{70863 + 5697 + 4949 + 10091} = 0.883777 \approx 0.884$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{10091}{10091 + 4949} = 0.670944 \approx 0.671$$

$$Specificity = \frac{TN}{TN + FP} = \frac{70863}{70863 + 5697} = 0.925588 \approx 0.926$$

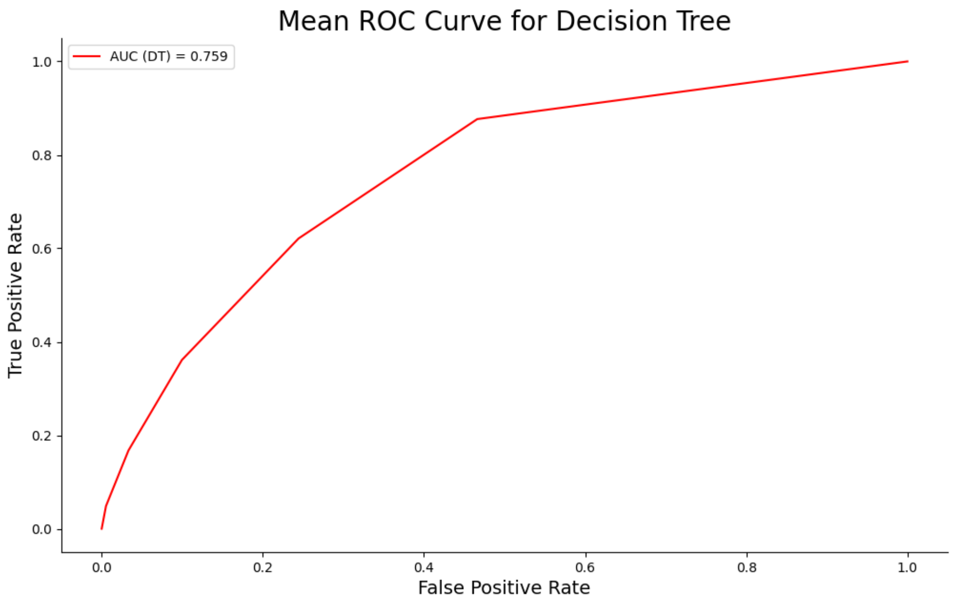
$$Precision = \frac{TP}{TP + FP} = \frac{10091}{10091 + 5697} = 0.639156 \approx 0.639$$

$$F_1 = \frac{2*(Precision*Sensitivity)}{Precision+Sensitivity} = \frac{2*(0.639156*0.670944)}{0.0.639156+0.670944} = 0.654655 \approx 0.655$$

Overall, the calculations show that the classification accuracy, sensitivity, specificity, precision, F_1 score of LR model are 0.884, 0.671, 0.926, 0.639, and 0.655, respectively. The accuracy value of 0.884 implies that the accuracy of the classification performance of DT model is 88.4%. In other words, the DT model’s classification accuracy for predicting students' enrollment decisions (i.e., enroll and non-enroll) is 88.4%. The sensitivity value of 0.671 implies that the probability of DT model to predict the students as enrolled when they actually enrolled is 67.1%. Similar to sensitivity, a specificity value of 0.926 shows that the model to predict students as non-enrolled when they actually did not enroll is 92.6%. The precision value of 0.639 implies the ratio of the actual number of enrolled students over the predicted number of enrolled students in the model. This means that 63.9% of students were actually enrolled in the predicted enrolled pool which the model predicted and classified students as enrolled. In addition to those four calculated values, the DT model's F_1 score value is 0.655. This is relatively lower than the accuracy (0.884) and specificity (0.926), which gives a hint that the DT model is less effective on showing predictive classification performance for imbalanced data. However, as mentioned earlier, there is no specific F_1 score threshold to determine the model’s classification performance efficiency. Therefore, mean PR-AUC was generated as an additional pertinent index to evaluate the model’s classification performance toward imbalanced data. This is presented after showing the mean ROC-AUC value and its evaluation.

After compiling five comparable metrics based on a concatenated confusion matrix, a mean ROC curve plot was created in Figure 16 to determine the AUC value. It shows that the average value of AUC is 0.759. After cross-matching the value with the AUC threshold for model performance evaluation, it falls between 0.7 and 0.8, giving insight that the DT model shows fair classification performance. However, this is an optimistic evaluation that disregards the data imbalance.

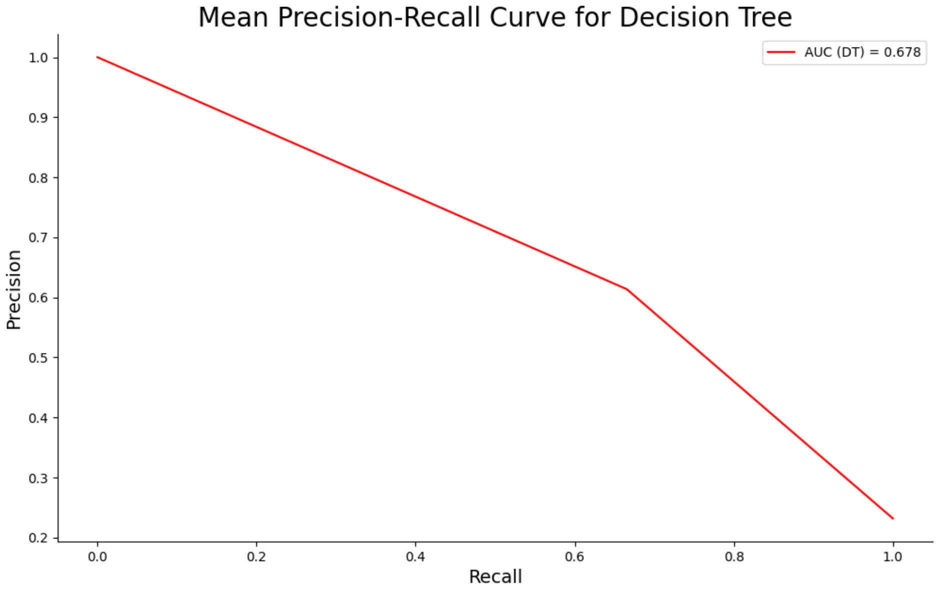
Figure 16. Mean ROC Curve for DT Model on Train/Validation Data with Stratified 10-fold Cross-Validation



Although the mean ROC curve plot and its AUC showed that the SVM model’s classification performance is good, Precision-Recall (PR) curve plot and related AUC were also compiled for a more rigorous evaluation of classification performance for imbalanced data. The mean PR curve for the SVM model was generated and shown in Figure 17. It shows that the value of

AUC for the mean PR curve of the DT model is 0.678. Since PR-AUC falls between 0.6 and 0.7 of the AUC threshold of the model's classification performance evaluation, the DT model seems to show poor classification performance on imbalanced data, specifically classifying the true positive cases. This gives a hint that the DT model is not efficient for making classification predictions on students as enrolled who were actually enrolled. Similar to how F_1 score was evaluated, PR-AUC gave a more detailed hint that the DT model was not efficient for making classification predictions on students as enrolled who actually enrolled.

Figure 17. Mean PR Curve for DT Model on Train/Validation Data with Stratified 10-fold Cross-Validation



Support Vector Machine (SVM)

As mentioned in the methodology chapter, the SVM model has three different types of kernels: linear, polynomial, and radial basis function. Since most of the college enrollment modeling starts based on a linear regression model (Solis, 2017; Perna & Titus, 2005; Klaauw, 2002), the current study compiled the SVM model using a linear kernel. To construct SVM with a linear kernel, determining the best cost parameter ('*C*') was needed to compile the SVM model with a linear kernel. However, there was no rule of thumb for choosing an optimistic value for *C*. Hence, the only option was to try a different value of *C* on the train and validation data to choose one that gives the lowest misclassification rate (Tay & Cao, 2002). According to Tay & Cao (2002), the general values used for a testing parameter for *C* range between 0.1 and 100. Hence, I applied the values of *C* with 0.1, 1, 10, and 100 to the SVM model with the computation of stratified 10-fold cross-validation. Therefore, 40 cases (4 × 10) of the SVM models were applied to the test and validation data for comparison. I used an svm() function to compile the SVM model and chose the one with the lowest misclassification rate (i.e., the highest accuracy rate). After testing multiple values of *C*, I decided to use an SVM model with *C*=10, which showed the highest average rates of sensitivity, precision and F_1 score on both the test and validation data.

Based on the selected SVM model with parameters of *C*=10, I created the concatenated confusion matrix and calculated the average values of five metrics; accuracy, sensitivity, specificity, precision, and F_1 score. The concatenated confusion matrix of the SVM model based on train and validation data is shown in Figure 18.

Figure 18. Concatenated Confusion Matrix of SVM Model on Train/Validation Data

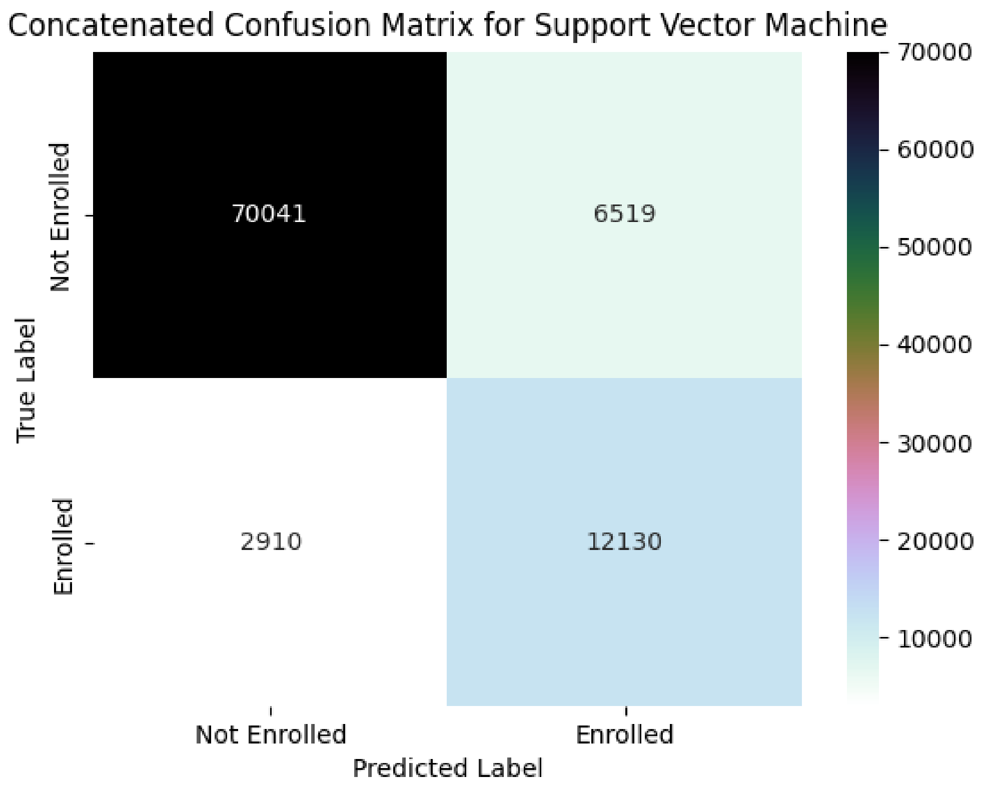


Figure 18 shows that the predicted number of enrolled students who actually enrolled (i.e., true positives) was 12,130, whereas the predicted number of enrolled students who actually did not enroll was 6,519 (i.e., false positives). Also, the predicted number of non-enrolled students who actually did not enroll was 70,041 (i.e., true negatives) and who actually enrolled was 2,910 (i.e., false negatives). Based on these outcomes, five evaluation indices of accuracy, sensitivity, specificity, precision, and F_1 score were calculated as below.

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} = \frac{70041 + 12130}{70041 + 6519 + 2910 + 12130} = 0.897063 \approx 0.897$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{12130}{12130 + 2910} = 0.806516 \approx 0.807$$

$$Specificity = \frac{TN}{TN + FP} = \frac{70041}{70041 + 6519} = 0.914851 \approx 0.915$$

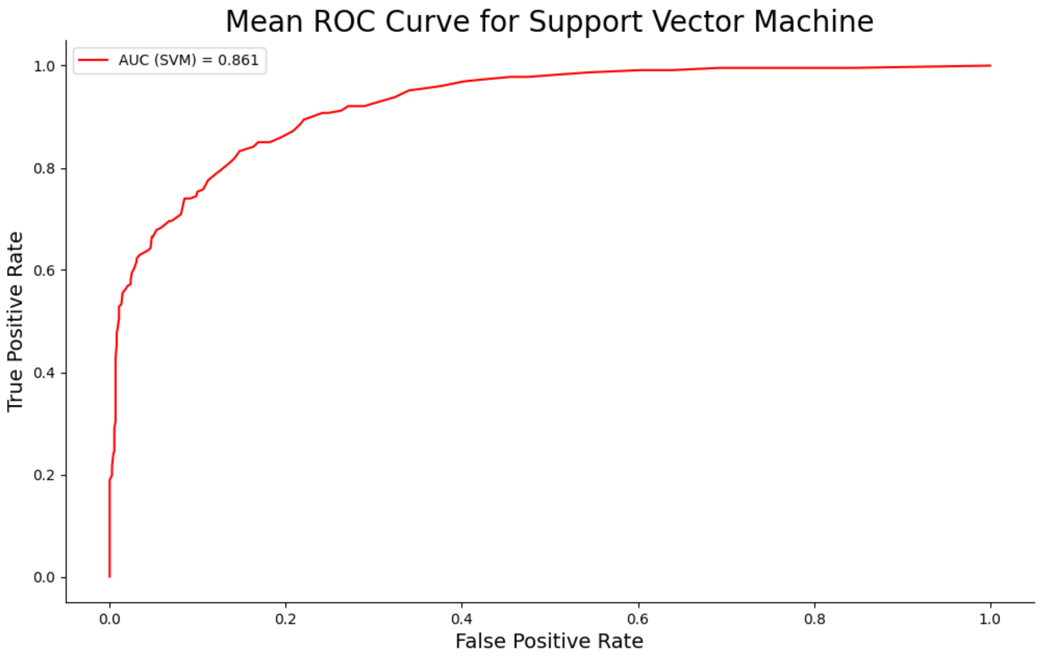
$$Precision = \frac{TP}{TP + FP} = \frac{12130}{12130 + 6519} = 0.650437 \approx 0.650$$

$$F_1 = \frac{2*(Precision*Sensitivity)}{Precision+Sensitivity} = \frac{2*(0.650437*0.806516)}{0.650437+0.806516} = 0.720116 \approx 0.720$$

The overall calculations show that the classification accuracy, sensitivity, specificity, precision, F_1 score of the SVM model are 0.897, 0.807, 0.915, 0.650, and 0.720, respectively. The accuracy value of 0.897 implies that the accuracy of the classification performance of the SVM model is 89.7%. In other words, the SVM model’s classification accuracy for predicting students' enrollment decisions (i.e., enroll and non-enroll) is 89.7%. The sensitivity value of 0.807 implies that the probability of the SVM model to predict the students as enrolled when they actually enrolled is 80.7%. Similar to sensitivity, a specificity value of 0.915 shows that the model to predict students as non-enrolled when they actually did not enroll is 91.5%. The precision value of 0.650 implies the ratio of the actual number of enrolled students over the predicted number of enrolled students in the model. This means that 65.0% of students were actually enrolled in the predicted enrolled pool which the model predicted and classified students as enrolled. In addition to those four calculated values, the F_1 score value of the SVM model is 0.720. This is relatively lower than the accuracy (0.897) and specificity (0.915), which gives a hint that the SVM model may be less effective in showing predictive classification performance for imbalanced data. However, as mentioned earlier, there is no specific threshold of F_1 score to determine the model’s classification performance. Hence, PR-AUC was generated as a salient index to evaluate the model’s classification performance toward imbalanced data. This is presented after showing the ROC-AUC value and its evaluation.

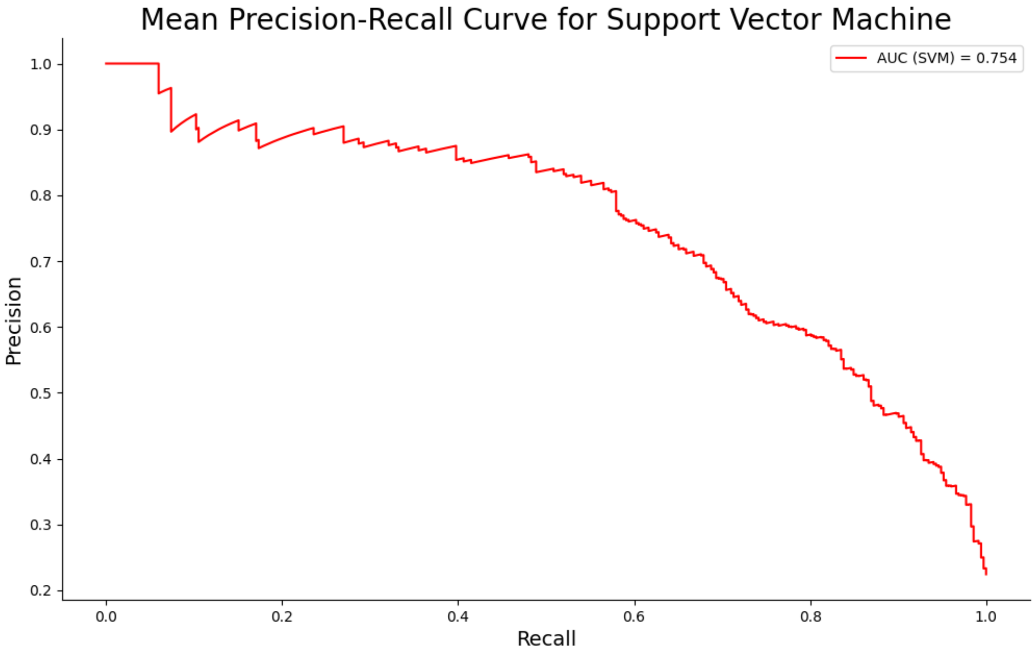
After compiling five comparable metrics based on a concatenated confusion matrix, a mean ROC curve plot was created in Figure 19 to determine the Area Under the Curve (AUC) value. It shows that the average value of AUC is 0.861, which falls between the AUC threshold of 0.8 and 0.9. Hence, based on the AUC value and its threshold, SVM was considered to have good classification performance. However, this is an optimistic evaluation that disregards the data imbalance.

Figure 19. Mean ROC Curve for SVM Model on Train/Validation Data with Stratified 10-fold Cross-Validation



Although ROC curve plot and its AUC provided an optimistic evaluation of SVM model's classification performance as good, Precision-Recall (PR) curve plot and related AUC was compiled for a more rigorous evaluation toward classification performance considering imbalanced data. The mean PR curve was generated and shown in Figure 20 showing that the value of AUC is 0.754. Since AUC falls between 0.7 and 0.8, SVM model seems to show fair classification performance on imbalanced data, specifically classifying the true positive cases. This gave a hint that SVM model was fairly efficient for making classification predictions on students as enrolled who were actually enrolled. Hence PR-AUC gave more detailed and sufficient insight than F_1 score which give indefinite insight that SVM model's classification is not effective in imbalanced data.

Figure 20. Mean PR Curve for SVM Model on Train/Validation Data with Stratified 10-fold Cross-Validation



Artificial Neural Network (ANN)

To compile the ANN model, it is important to determine the optimal numbers of the hidden layer(s) and node(s). As mentioned in the methodology chapter, the most common rules of thumb for choosing an optimal number of the hidden layer(s) and neuron(s) for ANN's decent performance are the following: 1) the number of hidden layers equals one, and 2) the number of neurons in that layer is the mean of the neurons in the input and output layers (Thomas et al., 2017). Hence on a hidden layer and 24 hidden nodes (i.e., (46 input nodes + 2 output nodes) / 2) were set up to run a model. Based on these settings, MLPClassifier() package was used to compile the ANN model and create concatenated confusion matrix. Based on the concatenated confusion matrix, I calculated the average values of five metrics of accuracy, sensitivity, specificity, precision, and F_1 score. Of course, the stratified 10-fold cross-validation method was applied, as well. The concatenated confusion matrix of the ANN model based on training and validated data is shown in Figure 21.

Figure 21. Concatenated Confusion Matrix of ANN Model on Train/Validation Data

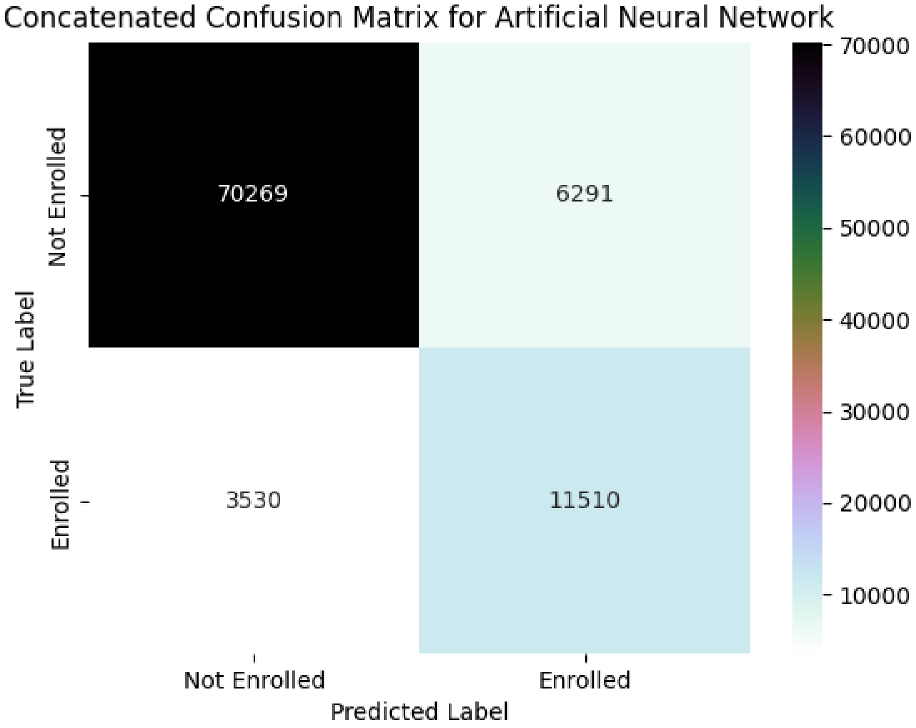


Figure 21 shows that the predicted number of enrolled students who actually enrolled (i.e., true positives) was 11,510, whereas the predicted number of enrolled students who actually were non-enrolled was 6,291 (i.e., false positives). Also, the predicted number of non-enrolled students who actually not enrolled was 70,269 (i.e., true negatives) and who actually enrolled was

3,530 (i.e., false negatives). Based these outcomes, five evaluation indices of accuracy, sensitivity, specificity, precision, and F_1 score were calculated below.

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} = \frac{70269 + 11510}{70269 + 6291 + 3530 + 11510} = 0.892784 \approx 0.893$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{11510}{11510 + 3530} = 0.765293 \approx 0.765$$

$$Specificity = \frac{TN}{TN + FP} = \frac{70269}{70269 + 6291} = 0.917829 \approx 0.918$$

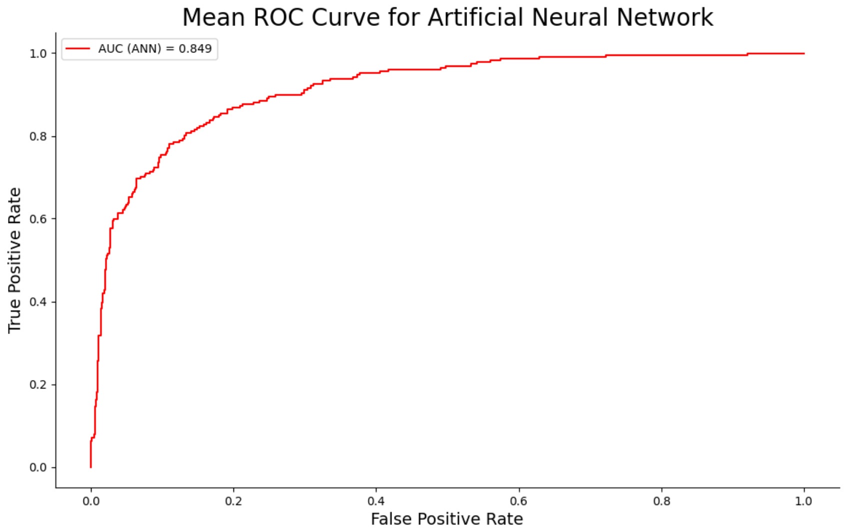
$$Precision = \frac{TP}{TP + FP} = \frac{11510}{11510 + 6291} = 0.646593 \approx 0.647$$

$$F_1 = \frac{2*(Precision*Sensitivity)}{Precision+Sensitivity} = \frac{2*(0.646593*0.765293)}{0.646593+0.765293} = 0.700953 \approx 0.701$$

Overall calculations show that the classification accuracy, sensitivity, specificity, precision, F_1 score of the ANN model are 0.893, 0.765, 0.918, 0.647, and 0.701, respectively. The accuracy value of 0.893 implies that the accuracy of the classification performance of the ANN model is 89.3%. In other words, the ANN model’s classification accuracy for predicting students' enrollment decisions (i.e., enroll vs. non-enroll) is 89.3%. The sensitivity value of 0.765 implies that the ANN model's probability of predicting the students as enrolled when they actually enrolled is 76.5%. Similar to sensitivity, a specificity value of 0.918 shows that the model to predict students as non-enrolled when they actually not enrolled is 91.8%. The precision value of 0.647 implies the ratio of the actual number of enrolled students over the predicted number of enrolled students of the model. This means that 64.7% of students were actually enrolled in the predicted enrolled pool which the model predicted and classified students as enrolled. In addition to those four calculated values, the ANN model's F_1 score value is 0.701. This is relatively lower than the accuracy (0.884) and specificity (0.926), which gives a hint that the ANN model is less effective in showing predictive classification performance for imbalanced data. However, there is no specific F_1 score threshold to determine the model's classification performance efficiency. Therefore, PR-AUC was generated as an additional pertinent index to evaluate the model's classification performance toward imbalanced data. This is presented after showing the ROC-AUC value and its evaluation.

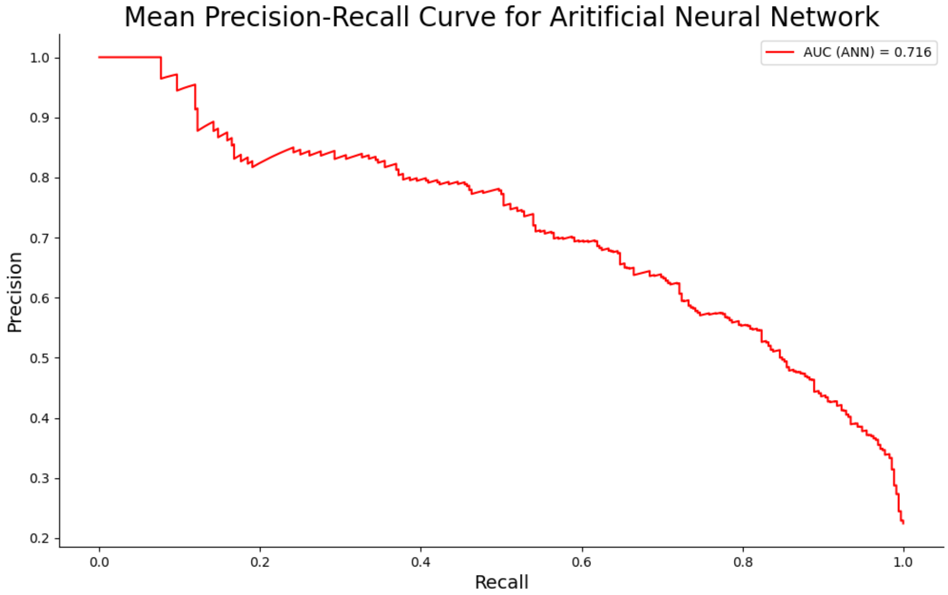
After compiling five comparable metrics based on a concatenated confusion matrix, a mean ROC curve plot was created to determine the Area Under the Curve (AUC) value. As mentioned in the methodology chapter, the AUC of ROC curve gave an idea of how well the model performs in compiling classification prediction. Since stratified 10-fold cross validation was applied, the mean of ten ROC curve plots were created and shown in Figure 22. It shows that the average value of AUC is 0.849. As mentioned earlier in the methodology chapter, the model with an AUC value between 0.8 and 0.9 were considered to have good classification performance. However, this is an optimistic evaluation that disregards the data imbalance.

Figure 22. Mean ROC Curve for ANN Model on Train/Validation Data with Stratified 10-fold Cross Validation



Although the ROC curve plot and its AUC evaluated that the ANN model shows a good classification performance, Precision-Recall (PR) curve plot and related AUC are also compiled for a more rigorous evaluation of classification performance for imbalanced data. Similar to how the ROC curve was generated, the mean PR curve was generated and shown in Figure 23. It shows that the value of AUC for the mean PR curve of the ANN model is 0.718. Since AUC falls between 0.7 and 0.8, the ANN model shows fair classification performance on imbalanced data, specifically classifying the true positive cases. This gave a detailed hint that the ANN model was fairly efficient for making classification predictions on students as enrolled who actually enrolled. Hence PR-AUC gave more detailed and sufficient insight than F_1 score, which gives indefinite insight that the ANN model's classification is ineffective in imbalanced data.

Figure 23. Mean PR Curve for ANN Model on Train/Validation Data with Stratified 10-fold Cross-Validation



Comparison of the Models

In this study, four models were trained to make predictions using a stratified 10-fold cross-validation method based on the validation data. Seven metrics were calculated to explain the performance of the models in making predictions based on validation data. As mentioned earlier, this study is more focused on identifying the students who will enroll in the next fall semester (i.e., class ‘1’ of the ‘Enrollment Decision’ variable), which was the true positive case of the models’ prediction. However, that does not imply that non-enrolled students are completely disregarded. As mentioned earlier, the study is also expected to provide insights into non-enrolled students for setting up complementary strategies for predicted non-enrolled students to enroll. Therefore, metrics (i.e., accuracy, specificity) that reflect true negative, false negative, and false positive cases were calculated and considered in addition to other metrics focused on true positive cases (i.e., sensitivity, precision).

Table 11 summarizes the four models' classification performance evaluation based on five metrics – accuracy, sensitivity, specificity, precision, F_1 score. It also summarized the models' classification performance evaluations that were made based on the ROC-AUC and PR-AUC values based on the AUC threshold. All four models showed a high classification rate in classifying non-enrolled students as non-enrolled (i.e., specificity; considering true negatives). This is natural since the data is highly imbalanced which the number of non-enrolled students is significantly greater than the number of enrolled students, with more cases to classify and predict. Moreover, the data imbalance also resulted in the models' high accuracy rates since the accuracy metric measured the models' overall classification performance in both true positive and true negative cases but was impacted by a significant amount of true negative cases (i.e., non-enrolled). Hence, for this study, these two metrics of specificity and accuracy reflected the classification performance of the data affiliated with the higher population (i.e., class '0' of the 'Enrollment Decision' variable). In other words, this study was generally expected to result in high classification accuracies and specificities because the model will predict a larger number of not enrolled students correctly compared to the enrolled students.

Considering the models' performance for enrolled students, metrics reflecting true positive cases are considered, which were sensitivity (i.e., recall), precision, and F_1 score. As mentioned earlier, the dataset is highly imbalanced, with a significantly small number of enrolled students relative to non-enrolled. Hence, it is highly likely that the models' sensitivity and precision rates will be lower than their accuracy and specificity rates. Comparing the sensitivity and precision and F_1 score, the SVM model turns out to show the best classification prediction performance on enrolled students (sensitivity = 0.806; precision = 0.650; F_1 score = 0.720) and the

ANN model turns out to be the second best (sensitivity = 0.765; precision = 0.645; F_1 score=0.701).

In addition to accuracy, sensitivity, specificity, precision, and F_1 score, ROC-AUC and PR-AUC values were considered for comparing the models' classification performance. As mentioned earlier, the ROC-AUC value is a general and robust indicator to evaluate the models' classification performance. Hence this metric was used to compare the classification performance across the models and SVM and ANN turned out to show good classification performance, whereas LR and DT showed fair classification performance. However, since ROC-AUC tends to give optimistic evaluation toward highly imbalanced data, PR-AUC was conducted to assess each model's classification performance toward imbalanced data, especially focused on classifying true positive cases (i.e., predicting students as enrolled who actually enrolled). This metric is also considered to be a more salient index than F_1 score, which gives indefinite insight into models' classification performance toward imbalanced datasets. Based on PR-AUC values, it turns out that SVM and ANN models showed fair predictive performance on classifying true positive cases, whereas LN and DT models were identified to show poor performance.

Overall, the SVM model had the highest accuracy (0.897), sensitivity (0.806), precision (0.650), F_1 score (0.720), ROC-AUC (0.861), and PR-AUC (0.754). Although it had the third lowest specificity (0.915) among the models, the SVM model still has a fairly high true negative rate. Hence, the SVM model is considered the best model showing good and fair classification performance for non-enrolled and enrolled students. Similar to SVM, the ANN model had the second-highest accuracy (0.893), sensitivity (0.765), precision (0.647), F_1 score (0.701), ROC-AUC (0.849), and PR-AUC (0.716). Although it had the second lowest specificity (0.918) across the models, the ANN model still has a fairly high true negative rate, like SVM. Hence, the ANN model is considered the second-best model showing good and fair classification performance for non-enrolled and enrolled students. Based on the above observations, the SVM and ANN were identified as the best algorithms to implement for compiling college enrollment prediction models for the test dataset.

Table 12. Comparison of Metrics from Different Models on the Train/Validation Data

	Accuracy	Sensitivity	Specificity	Precision	F_1 Score	ROC-AUC	ROC-AUC Evaluation	PR-AUC	PR-AUC Evaluation
LR Model	0.861	0.674	0.897	0.564	0.614	0.767	Fair	0.639	Poor
DT Model	0.884	0.671	0.926	0.639	0.655	0.759	Fair	0.661	Poor
SVM Model	0.897	0.807	0.915	0.650	0.720	0.861	Good	0.754	Fair
ANN Model	0.893	0.765	0.918	0.647	0.701	0.849	Good	0.716	Fair

Selected Models Application on Test Data

Based on the compiled four machine learning models using train and validation data and its comparison, SVM and ANN algorithm models were chosen because of their high values of accuracy, sensitivity, precision, F_1 score, ROC-AUC, and PR-AUC metrics. These two models were applied using test data, which reflected the students' admission data for Fall 2019. For the test dataset, a total of 17,198 admits, including 2,636 enrolled (15%) and 14,562 not enrolled (85%) students, were included. Stratified 10-fold cross-validation was not applied since it was already used for training/validation data to reduce the probability of over/underfitting and bi-asedness.

Applying Support Vector Machine Model

As mentioned earlier, the SVM model with *C*=10 with gamma=1 was implemented to the test data and the confusion matrix was compiled. Figure 20 shows the confusion matrix for test data of the Fall 2019 cohort.

Figure 24. Confusion Matrix of SVM Model on Test Data

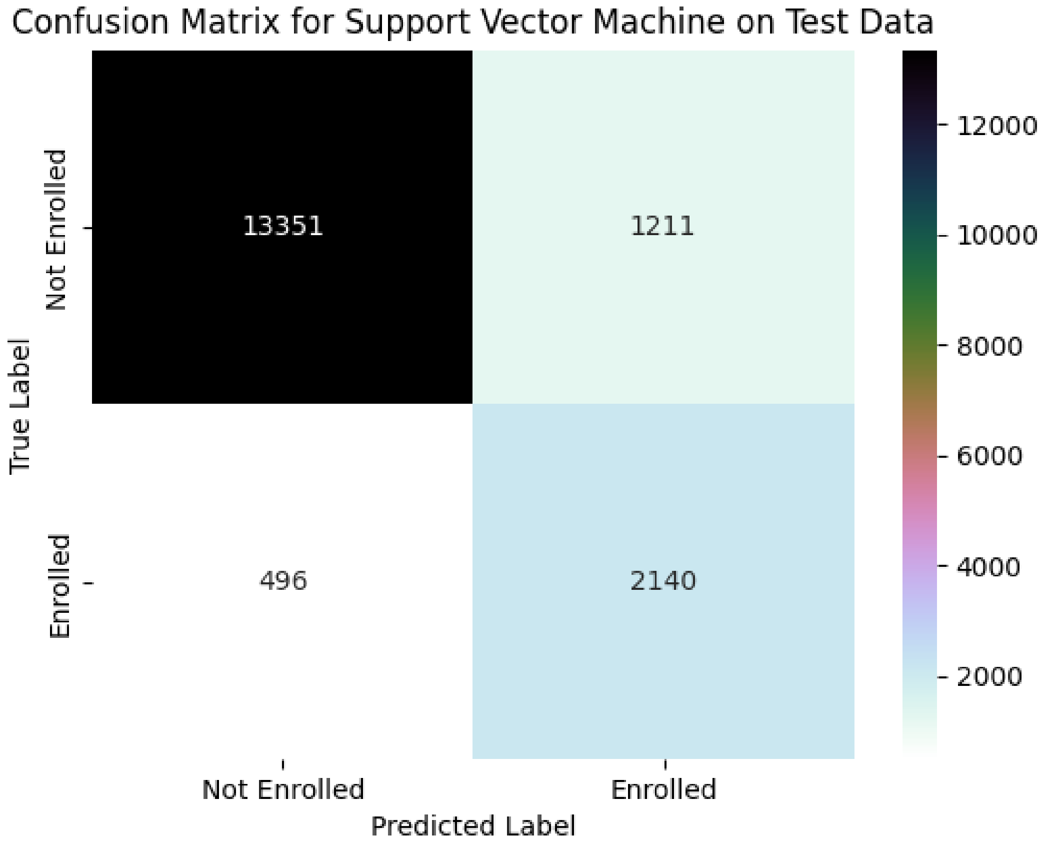


Figure 24 shows that the predicted number of enrolled students who actually enrolled (i.e., true positives) was 2,140 whereas the predicted number of enrolled students who actually were non-enrolled was 1,211 (i.e., false positives). Also, the predicted number of non-enrolled students who actually did not enrolled was 13,351 (i.e., true negatives) and who actually enrolled was 496 (i.e., false negatives). Based these outcomes, five evaluation indices of accuracy, sensitivity, specificity, precision, and F_1 score were calculated below.

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} = \frac{13351 + 2140}{13351 + 1211 + 496 + 2140} = 0.900744 \approx 0.901$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{2140}{13351 + 496} = 0.811836 \approx 0.812$$

$$Specificity = \frac{TN}{TN + FP} = \frac{13351}{13351 + 1211} = 0.916838 \approx 0.917$$

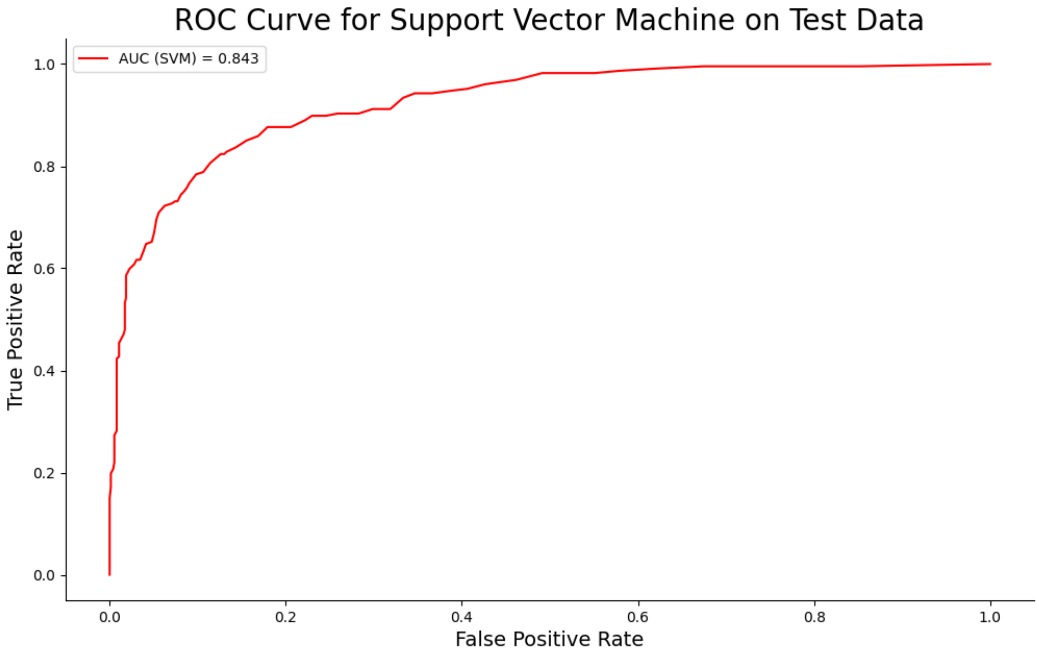
$$Precision = \frac{TP}{TP + FP} = \frac{2140}{2140 + 1211} = 0.638615 \approx 0.639$$

$$F_1 = \frac{2*(Precision*Sensitivity)}{Precision+Sensitivity} = \frac{2*(0.638615*0.811836)}{0.638615+0.811836} = 0.714882 \approx 0.715$$

Overall calculations show that the classification accuracy, sensitivity, specificity, precision, F_1 score of the SVM model are 0.901, 0.812, 0.917, 0.639, and 0.715, respectively. The accuracy value of 0.888 implies that the accuracy of the classification performance of the ANN model is 90.1%. In other words, the SVM model's classification accuracy for predicting students' enrollment decisions (i.e., enroll vs. non-enroll) is 90.1%. The sensitivity value of 0.812 implies that the probability of the SVM model to predict the students as enrolled when they actually enrolled is 81.2%. Similar to sensitivity, a specificity value of 0.917 shows that the model to predict students as non-enrolled when they actually not enrolled is 91.7%. In addition to those four calculated values, the SVM model's F_1 score value is 0.715. This is relatively lower than the accuracy (0.901) and specificity (0.917), which gives a hint that the SVM model is less effective on showing predictive classification performance for imbalanced data. However, since there is no specific F_1 score threshold to determine the model's classification performance efficiency, PR-AUC was generated as an additional pertinent index to evaluate the model's classification performance toward imbalanced data. This is presented after showing the ROC-AUC value and its evaluation.

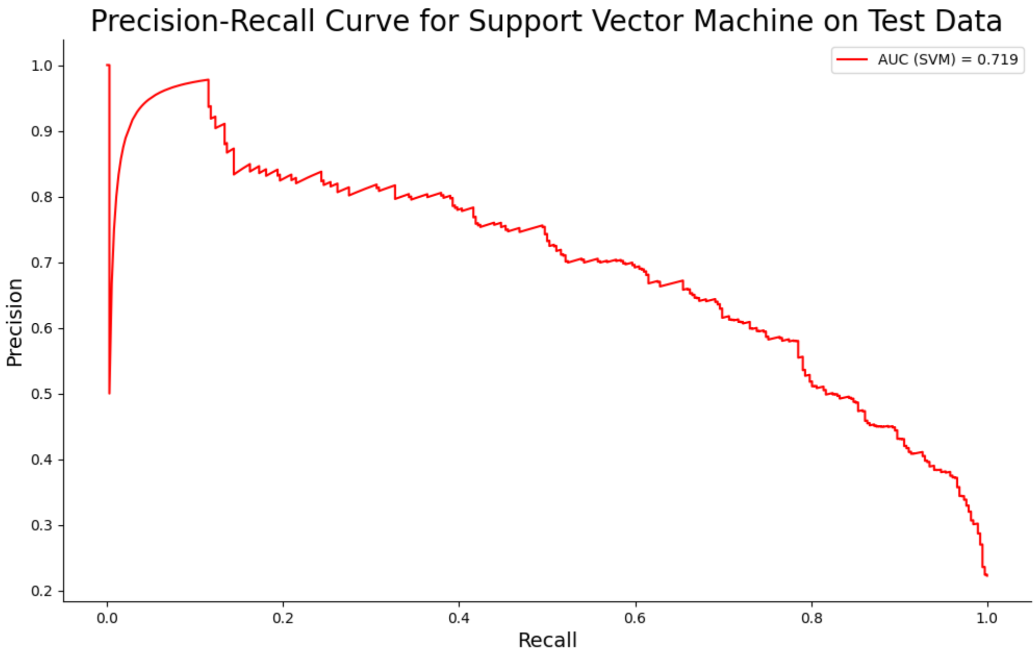
After compiling five comparable metrics based on a confusion matrix, a ROC curve plot was created to determine the Area Under the Curve (AUC) value. As mentioned in the methodology chapter, the AUC of ROC curve is a general index of determining how well the model performs in compiling classification prediction. Hence, the ROC curve plots were created and shown in Figure 25. It shows that the AUC value of the SVM model is 0.843. Since this value falls between 0.8 and 0.9 of the AUC threshold, the SVM model was considered to have good classification performance. However, ROC-AUC tends to provide an optimistic evaluation of the model even though the data is highly imbalanced. Therefore, the PR-AUC metric was also generated to evaluate the model's classification performance more rigorously.

Figure 25. ROC Curve for SVM Model on Test Data



Although the ROC curve plot and its AUC evaluated SVM model as an efficient classifier, Precision-Recall (PR) curve plot and related AUC were also compiled for more rigorous evaluation of classification performance for imbalanced data. Similar to how the ROC curve was generated, the mean PR curve was generated and shown in Figure 26. It shows that the value of AUC for mean PR curve of the SVM model is 0.719. Since the AUC value falls between 0.7 and 0.8 AUC threshold, the SVM model seems to show fair classification performance on imbalanced data, specifically classifying the true positive cases. This gave a hint that the SVM model was efficient for making classification predictions on students as enrolled who were actually enrolled for test data.

Figure 26. PR Curve for SVM Model on Test Data



Overall, the SVM model's performance on test data shows that all seven-evaluation metrics are close enough to the one I achieved on train/validation data. Hence, this gives an insight that the model is actually learning from the training/validation data and can generalize as long as the new incoming data is applied with the identical data structure. Moreover, it provides another insight into the reliability and uniformity of the model predictions on the combined cohort year data and individual level. Based on the above observations, I reconfirm and conclude that the SVM is a good prediction model for the dataset.

Applying Artificial Neural Network Model

As mentioned earlier, the ANN model with one hidden layer and 25 hidden nodes were compiled for test data application. Based on these settings, I created a confusion matrix to calculate the average values of five metrics of accuracy, sensitivity, specificity, precision, and F_1 score. The confusion matrix of the ANN model based on test data is shown in Figure 26.

Figure 27. Confusion Matrix of ANN Model on Test Data

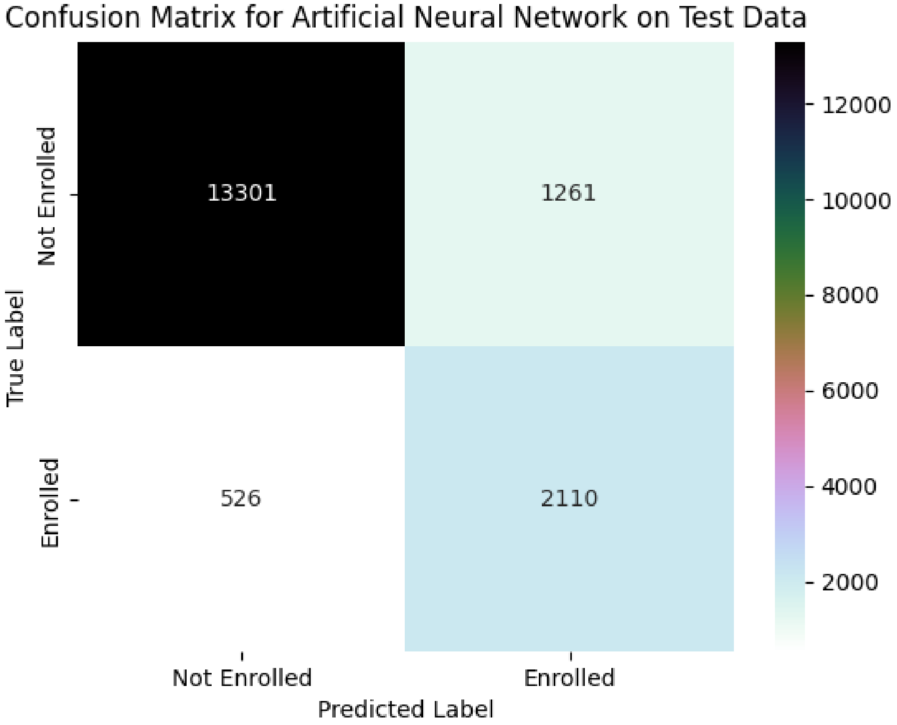


Figure 27 shows that the predicted number of enrolled students who actually enrolled (i.e., true positives) was 2,110 whereas the predicted number of enrolled students who actually were non-enrolled was 1,261 (i.e., false positives). Also, the predicted number of non-enrolled students who were actually not enrolled was 13,301 (i.e., true negatives) and who actually enrolled was 526 (i.e., false negatives). Based on these outcomes, five evaluation indices of accuracy, sensitivity, specificity, precision, and F_1 score were calculated below.

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} = \frac{13301 + 2110}{13301 + 1261 + 526 + 2110} = 0.896093 \approx 0.896$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{2110}{13301 + 526} = 0.800455 \approx 0.800$$

$$Specificity = \frac{TN}{TN + FP} = \frac{13301}{13301 + 1261} = 0.913405 \approx 0.913$$

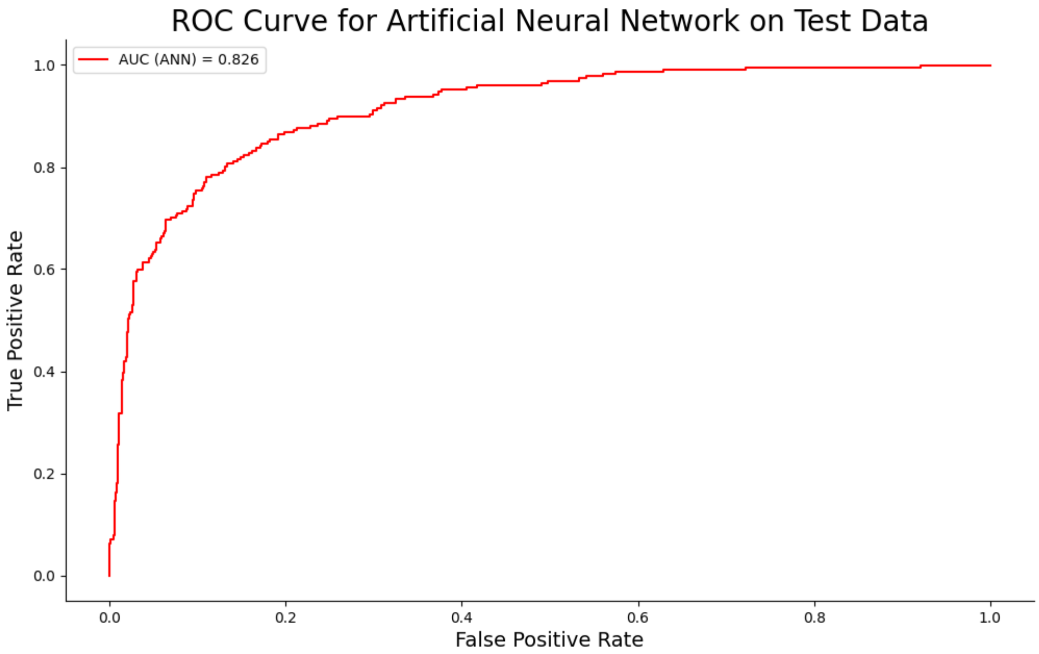
$$Precision = \frac{TP}{TP + FP} = \frac{2110}{2110 + 1261} = 0.625927 \approx 0.626$$

$$F_1 = \frac{2*(Precision*Sensitivity)}{Precision+Sensitivity} = \frac{2*(0.625927*0.800455)}{0.625927+0.800455} = 0.702514 \approx 0.702$$

Overall calculations show that the classification accuracy, sensitivity, specificity, precision, F_1 score of the SVM model are 0.896, 0.800, 0.913, 0.626, and 0.702, respectively. The accuracy value of 0.896 implies that the accuracy of the classification performance of the ANN model is 89.6%. In other words, the ANN model’s classification accuracy for predicting students' enrollment decisions (i.e., enroll vs. non-enroll) is 89.6%. The sensitivity value of 0.800 implies that the ANN model's probability of predicting the students as enrolled when they actually enrolled is 80.0%. Similar to sensitivity, a specificity value of 0.913 shows that the model to predict students as non-enrolled when they are actually not enrolled is 91.3%. The precision value of 0.626 implies the ratio of the actual number of enrolled students over the predicted number of enrolled students of the model. This means that 62.6% of students were actually enrolled in the predicted enrolled pool which the model predicted and classified students as enrolled. In addition to those four calculated values, the ANN model's F_1 score value is 0.702. This is relatively lower than the accuracy (0.895) and specificity (0.913), which gives a hint that the ANN model is less effective on showing predictive classification performance for imbalanced data. Since there is no specific F_1 score threshold to determine the model’s classification performance efficiency, PR-AUC was generated as an additional pertinent index to evaluate the model’s classification performance toward imbalanced data. This is presented after showing the ROC-AUC value and its evaluation.

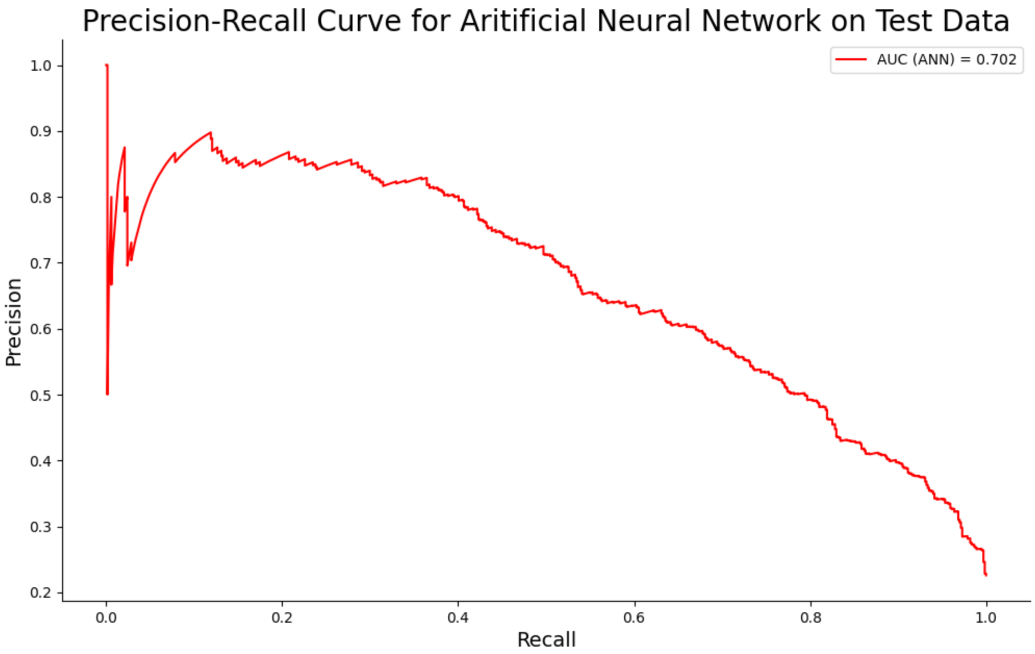
After compiling five comparable metrics based on a confusion matrix, a ROC curve plot was created in Figure 28 to determine the Area Under the Curve (AUC) value. It shows that the value of AUC is 0.826, which indicates that the ANN model is showing a good classification prediction performance because it falls between 0.8 and 0.9 of the AUC value threshold. However, the ROC-AUC value may provide an optimistic evaluation while disregarding the data imbalance.

Figure 28. ROC Curve for ANN Model on Test Data



Since the ROC curve plot and its AUC provide an optimistic evaluation of highly imbalanced data, the Precision-Recall (PR) curve plot and related AUC were compiled for a more rigorous assessment of classification performance for imbalanced data. Similar to how the ROC curve was generated, the PR curve was generated and shown in Figure 29. It shows that the value of AUC for mean PR curve of the ANN model is 0.702. Since AUC falls between 0.7 and 0.8, the ANN model shows fair classification performance on imbalanced data, specifically classifying the true positive cases. This gave a hint that the ANN model was fairly efficient for making classification predictions on students as enrolled who were actually enrolled.

Figure 29. PR Curve for ANN Model on Test Data



Overall, the ANN model's performance on test data shows that all seven evaluation metrics are close enough to the one I achieved on train/validation data. Hence, this also gives the same insight as the SVM model that the ANN model is actually learning from the training/validation data and can generalize as long as the new incoming data is applied with the identical data structure. Moreover, it gives another insight into the reliability and uniformity of the model predictions on the combined cohort year data and individual level. Based on the above observations, I reconfirm and conclude that the ANN is a good prediction algorithm model for the dataset along with the SVM algorithm model.

Feature Importance on Selected Models

After comparing the models and identifying the SVM and the ANN models as the best models in classification performance, these models were used to determine the most significant factors that contributed to predicting the student's enrollment decisions, particularly as "enrolled." Based on the insights of the reliability and uniformity of the model predictions on the combined cohort year data, these two models were applied to train/validation and test data combined and calculated the significant factors. The SVM and ANN algorithm models were generated by compiling the drop-column importance method to see how much each independent variable affects the model's classification performance.

The lists of 46 factors of SVM and ANN are shown in Tables 13 and 14. Those are ordered from the most important to the least important. These feature importances are normalized and add up to 1. By normalizing the data, the features have the same order of magnitude and scatter, making it easier to find which is more relevant. Based on the feature importance findings of two models, SVM and ANN, the magnitude of each factor influencing outcomes differ across the models. However, the top five factors that were identified as most important were the same; Distance from Campus (i.e., proximity), High School GPA, Expected Family Contribution, Unemployment Rate, and Institutional Financial Aid Total amount. This gives a hint that these top five factors are the ones that impact students' enrollment decisions significantly relative to other rest of the 41 factors. On the other hand, the factors with low importance values have less impact on the model's performance (i.e., enrollment decisions).

Table 13. List of Factors by Feature Importance using SVM Model

Rank	Attributes	Feature Importance
1	Distance_from_Campus	0.289942
2	Unemployment_Rate	0.185686
3	HSGPA	0.035956
4	Fed_Efc	0.025896
5	Inst_Fin_Aid_Total	0.023618
6	SAT_ACT_SuperScore	0.023113
7	Inflation_Rate	0.021373
8	PEL_Graduate_Degree	0.021133
9	HSCluster_70	0.018799
10	Sex_M	0.016743
11	HSCluster_68	0.016156
12	PEL_Some_College	0.016112
13	Race_Asian	0.016073
14	PEL_High_School_Graduate	0.015902
15	HSCluster_65	0.015134
16	HSCluster_67	0.014869
17	Race_Hispanic	0.014524
18	Race_Other	0.014268
19	PEL_Associates_Degree	0.013683
20	HSCluster_75	0.012576
21	Race_Black_AA	0.012116
22	PEL_Some_High_School	0.012076
23	HSCluster_73	0.012049
24	HSCluster_72	0.011787
25	Race_Unknown	0.011599
26	HSCluster_59	0.011549
27	HSCluster_57	0.011400
28	HSCluster_62	0.010904
29	HSCluster_66	0.010889
30	HSCluster_76	0.010813
31	HSCluster_64	0.010572
32	HSCluster_78	0.010533
33	HSCluster_58	0.007757
34	HSCluster_55	0.007663
35	HSCluster_52	0.007592
36	HSCluster_53	0.006557
37	HSCluster_74	0.005500
38	HSCluster_63	0.005168
39	HSCluster_56	0.005162
40	HSCluster_60	0.003882
41	HSCluster_77	0.002876
42	HSCluster_69	0.000000
43	HSCluster_54	0.000000
44	HSCluster_61	0.000000
45	HSCluster_51	0.000000
46	HSCluster_71	0.000000

Table 14. List of Factors by Feature Importance using ANN Model

Rank	Attributes	Feature Importance
1	Distance_from_Campus	0.162213
2	HSGPA	0.158321
3	Fed_Efc	0.140134
4	Unemployment_Rate	0.103597
5	Inst_Fin_Aid_Total	0.100865
6	Inflation_Rate	0.096071
7	SAT_ACT_SuperScore	0.075527
8	HSCluster_70	0.017555
9	HSCluster_68	0.016027
10	PEL_Graduate_Degree	0.015676
11	Sex_M	0.012766
12	Race_Asian	0.011323
13	Race_Hispanic	0.010383
14	PEL_High_School_Graduate	0.009212
15	PEL_Some_College	0.008333
16	Race_Other	0.007823
17	HSCluster_65	0.007343
18	HSCluster_67	0.007154
19	PEL_Associates_Degree	0.005710
20	Race_Black_AA	0.005520
21	HSCluster_75	0.005165
22	PEL_Some_High_School	0.004421
23	HSCluster_72	0.002806
24	HSCluster_73	0.002293
25	HSCluster_59	0.001956
26	Race_Unknown	0.001692
27	HSCluster_62	0.001535
28	HSCluster_66	0.001214
29	HSCluster_76	0.001138
30	HSCluster_64	0.001123
31	HSCluster_57	0.000667
32	HSCluster_58	0.000653
33	HSCluster_78	0.000623
34	HSCluster_53	0.000604
35	HSCluster_60	0.000601
36	HSCluster_74	0.000486
37	HSCluster_63	0.000379
38	HSCluster_56	0.000367
39	HSCluster_52	0.000337
40	HSCluster_55	0.000234
41	HSCluster_69	0.000078
42	HSCluster_71	0.000077
43	HSCluster_77	0.000000
44	HSCluster_61	0.000000
45	HSCluster_54	0.000000
46	HSCluster_51	0.000000

Potential Predictive Probability for Students to Enroll

In addition to the feature importance analysis, the predictive probability of students to enroll or not was calculated at an individual level for both models, SVM and ANN. In general, machine learning classifiers don’t just give binary predictions but provide numerical values between 0 and 1 for their predictions. These numbers are sometimes called the model score or confidence regarding each binary decision (i.e., enroll vs. not enroll). It is a way for the model to express its certainty about what class the input data belongs to. In most applications, the exact score is ignored and use a threshold to round the score to a binary answer as enroll or not enroll. Calibration transforms these scores into probabilities and is used more effectively in decision-making.

The SVM's methodology focuses on finding the best linear classifier(s) that can classify two different outcomes. It also identifies the most impactful features (i.e., independent variables) on the outcomes (i.e., dependent variable). However, this method does not compute the probability of affiliation to each group, which is one of the relative weaknesses of SVM compared with LR and ANN. To overcome this issue, many researchers have proposed transforming SVM to calculate the estimated probability (Platt et al., 2000; Sollich, 2002). Hence, in this study, the SVM model was transformed slightly based on a proposal by Platt et al. (2000). This method calculates the conditional posterior probability by measuring the distance between the data and the classifier. Table 15 shows the output of the top ten student records with the highest probability of being classified into the enrolled group. Because the original data is sourced as de-identified, only each student’s influential enrollment factors were listed along with the probability.

For the ANN model, input variables are received by the input layer, and then the hidden layer performs predicted probability computations based on these input variables. As previously mentioned, this study used a single hidden layer in the ANN model, and a sigmoid function was used as the activation function in the output layer. Since the sigmoid function's output was between 0 and 1, the ANN model generated values between 0 and 1. The numbers between 0 and 1 imply the probability of students' decision to enroll. Hence, a higher probability value indicated that the students were more likely to enroll. The records of students with the top 10 highest enrollment probabilities are shown in Table 16.

Table 15. Top 10 Student Records with the Highest Probabilities of Making Enrollment Decisions as “Enrolled” using SVM Model

SAT_ACT_	Race_	Race_		Distance_from_	Unemployment_		Inst_Fin_	Enroll_	Predicted	Predicted
SuperScore	Black_AA	Hispanic	…	Campus (Miles)	Rate (%)	…	Aid_Total ($)	Decision	Probabilitty Enroll	Probabilitty Not Enroll
31	0	0		28.8	4.9		10500	1	0.998730	0.001270
19	1	0		353.5	4.9		43186	1	0.998671	0.001329
32	0	0	…	245.5	4.9	…	8000	1	0.998556	0.001444
30	0	0		259.7	4.9		8000	1	0.998339	0.001661
29	0	0		35.0	4.9		9500	1	0.997907	0.002093
33	0	0		353.3	4.9		41000	1	0.997877	0.002123
31	0	0	…	332.4	4.9	…	10500	1	0.997720	0.002280
24	1	0		26.7	4.9		51286	1	0.997581	0.002419
33	1	0		27.7	4.9		40700	1	0.997484	0.002516
25	0	0		4.9	7.6		6246	1	0.997221	0.002779

Table 16. Top 10 Student Records with the Highest Probabilities of Making Enrollment Decisions as “Enrolled” using ANN Model

SAT_ACT_	Race_	Race_		Distance_from_	Unemployment_		Inst_Fin_	Enroll_	Predicted	Predicted
SuperScore	Black_AA	Hispanic	…	Campus (Miles)	Rate (%)	…	Aid_Total ($)	Decision	Probabilitty Enroll	Probabilitty Not Enroll
24	0	0		12.6	4.0		76762	1	0.986383	0.013617
32	0	0		422.4	4.0		63653	1	0.974848	0.025152
25	0	1	…	923.4	4.0	…	60753	1	0.973759	0.026241
23	1	0		960.7	4.5		59018	1	0.971441	0.028559
23	0	0		6.5	4.0		62910	0	0.971155	0.028845
30	0	0		2.4	4.9		62231	1	0.970045	0.029955
30	0	0	…	433.2	4.5	…	60258	1	0.969829	0.030171
22	0	1		910.0	3.7		57358	1	0.968384	0.031616
29	0	0		170.1	4.0		60408	1	0.968196	0.031804
25	0	0		2.4	5.4		61095	0	0.968144	0.031856

CHAPTER FIVE

DISCUSSIONS

The primary objective of this study was to conduct a comparative analysis of four machine learning models on students' college enrollment decision predictions and identify the algorithm model, showing the best classification performance. Moreover, another objective was to show how the selected models were implemented for conducting feature importance and predictive probability analyses, so the institutions could get insights on how to apply these models to the data for pragmatic use. The research design, data collection, and analysis of this study is well-tied with Perna's (2006) conceptual framework, which provided a comprehensive set of concepts and ideas to understand and explain college enrollment decisions. Specifically, this framework informed the selection and examination of the key variables, which brough the practical perspective when investigating the statistical methodology. In other words, by utilizing Perna's framework, the study ensured its research methodology was relevant, rigorous, and focused, thereby contributing to an existing body of knowledge on higher education enrollment management.

The following sections discuss the major topics in detail, including the proposed three research questions outcomes, data imbalance, limitations, implementations, and future studies.

Comparative Analysis of Four Machine Learning Algorithms on Train/Validation Data

As mentioned in the literature review chapter, a few studies conducted U.S. college enrollment prediction modeling for students likely to enroll using machine learning algorithms.

These studies conducted comparative analysis determining which machine learning algorithms showed the best prediction in performance. These comparative analyses of machine learning algorithm applications were broadly classified into the following categories.

- ANN vs. SVM (Walczk & Sincich, 1999; DesJardins & Gonzales, 2002; Gerasimovic & Bugaric, 2018)
- LR vs. ANN vs. DT (Antons & Maltz, 2006; Chang, 2006)
- DT vs. KNN vs. NB (Vialardi et al., 2011)
- ANN vs. DT vs. SVM vs. KNN vs. RF (Ragab et al., 2014)
- ANN vs. SVM (Lux et al., 2016)
- LR vs. NN vs. BN vs. RF vs. SVM (Cirelli et al., 2018)
- LR vs. SVM (Slim et al., 2019)
- LR vs. NB vs. DT vs. SVM vs. KNN vs. RF vs. GB (Basu et al., 2019)

The best prediction models selected by previous studies were LR, ANN, DT, and SVM. However, no research conducted a comparative analysis using those four machine learning algorithms that were identified as the best. This study presented the comparative analysis for those four machine learning models regarding college enrollment prediction performance primarily focused on students likely to enroll. Hence, this is one of the contexts in which my study differs from the above research. The following subsection discusses each algorithm in detail.

Logistic Regression (LR) Algorithm Implementation

LR model has been a general classification algorithm for compiling a college enrollment prediction model. As presented in the result chapter, it turns out that the logistic regression has higher accuracy (0.861), specificity (0.897), and ROC-AUC (0.767) rates relative to sensitivity

(0.674), precision (0.564), F_1 score (0.614), and PR-AUC (0.639). This implies that the model is more prone to classify and predict non-enrolled students as non-enrolled than predicting enrolled students as enrolled. In other words, the accuracy and specificity rates are significantly higher because the size of the non-enrolled students' pool is significantly greater than the enrolled pool in which the data are more likely to capture the non-enrolled students and predict as non-enrolled than capturing enrolled students and predict them as enrolled.

As mentioned in the literature review, the LR model has been the most popular empirical model for college enrollment since the 1980s. Since there were fewer college admission data with less number of influential factors to consider that impact students' enrollment decisions, the LR model was an appropriate model to apply. However, today, the size of the college admission data is large and abundant, with various influential factors on college enrollment. Hence, many higher education institutions are confronting implementing efficient methodologies for handling big and high-dimensional data. However, the LR model is prone to overfitting in high dimensional data because those data may involve complex relationships across predictor (i.e., input) variables and outcome variables (Dreiseitl & Ohno-Machado, 2002). Also, LR can be more sensitive to outliers, leading to a decision boundary that does not generalize well to new data (Hosmer, 2013). Therefore, the LR's classification prediction performance was relatively ineffective compared to DT, SVM and ANN. That is because DT, SVM and ANN are more robust to outliers than LR (Sakr et al., 2017). Since DT can handle outliers by splitting the data into subsets based on the input features, it can help avoid overfitting outliers (Herzog, 2006; Gomex & Almeida, 2017). In addition, SVM seeks to maximize the margin between the decision boundary and the closest data points, which can help to avoid overfitting to outliers (Hastie et al., 2004;

Platt, 2000; Hansen & Sargent, 2001). In addition, SVM and ANN models are more flexible and better at adjusting to the data with a more dynamic approach for handling high-dimensional data (Hastie et al., 2004; Hansen & Sargent, 2001; Dreiseitl & Ohno-Machado, 2002). For example, the best SVM model is compiled by attempting to apply multiple cost parameters, *C*, and choose the optimal one among them, tuning its' model to show the best classification prediction performance. Therefore, it is less prone to overfitting than LR. Like SVM, the ANN model is compiled by selecting the optimal number of hidden layers and nodes, making its model the best classification prediction performance. Although there is a basic rule of thumb to determine the number of hidden layers and nodes, multiple numbers of hidden layers and nodes can be attempted to make the best choice of those values, making the ANN model perform better on classification prediction compared to LR.

Decision Tree (DT) Algorithm Implementation

DT model has been a general classification algorithm for compiling a college enrollment prediction model. However, this model was more computationally expensive compared to LR. In other words, compiling a model to program in Python took longer than LR.

Similar to LR, DT had higher accuracy (0.884), specificity (0.926), and ROC-AUC (0.759) rates relative to sensitivity (0.671), precision (0.639), F_1 score (0.655), and PR-AUC (0.661). This implies that the DT model is also more prone to classify and predict non-enrolled students as non-enrolled than predicting enrolled students as enrolled. In other words, the accuracy and specificity rates are higher because the size of the non-enrolled students' pool is greater than the enrolled pool, in which the data are more likely to capture the non-enrolled students and predict as non-enrolled compared to capturing enrolled students and predict them as enrolled.

Although DT is also considered a popular model to apply with providing easier interpretations, the current study showed that it’s classification prediction performance was ineffective compared to SVM and ANN. That is because DT algorithms are more unstable to apply to high-dimensional data relative to SVM and ANN, which may easily result in the under/overfitting of the data. Also, as mentioned earlier, SVM and ANN models are more flexible and better than DT at adjusting to the data with a more dynamic approach for handling high-dimensional data (Hastie et al., 2004; Hansen & Sargent, 2001; Dreiseitl & Ohno-Machado, 2002). Moreover, the SVM and ANN models are more robust to noise data than DT because linear SVM seeks to find a decision boundary that maximizes the margin between different classes (Hastie et al., 2004; Platt, 2000; Hansen & Sargent, 2001). In contrast, DT can be relatively sensitive to noise and outliers, which can lead to overfitting or inaccurate predictions (Herzog, 2006; Gomex & Almeida, 2017). Also, DT can create complex decision boundaries that overfit the training data, whereas ANN can learn to generalize from noisy data, which can lead to better performance on unseen data (Herzog, 2006).

Support Vector Machine (SVM) Algorithm Implementation

As mentioned in the results chapter, a linear SVM with a cost parameter value of 10 was implemented in the train/validation dataset. The results showed that the SVM had high accuracy (0.897), specificity (0.915), and ROC-AUC (0.861) rates relative to sensitivity (0.807), precision (0.650), F_1 score (0.720), and PR-AUC (0.754). This may be reflected as having similar results to LR, DT, and ANN. However, it turns out that the SVM model is classifying imbalanced data fairly well, which has the highest sensitivity, precision, F_1 score and PR-AUC values compared to other models. In other words, the SVM model shows a better classification prediction performance for enrolled students as enrolled.

As mentioned earlier, one of the advantages of the SVM algorithm is that it is more robust than LR and DT in handling multiple feature spaces (Auria & Moro, 2008). In other words, SVM was more prone to handling complex, high-dimensional, and imbalanced data by applying an optimal cost parameter value. Hence, it had less risk of overfitting on the validation and test dataset. Moreover, the SVM model is less sensitive to noisy data, such as including outliers (Sakr et al., 2017). Hence it is relatively well suited to making good predictions compared to LR and DT (Pochet & Suykens, 2006). Also, linear SVM seeks to maximize the margin between different classes, which can lead to a decision boundary that generalize well to new data. In contrast, ANN can be more prone to overfitting due to their flexibility and the potential for the model to memorize noise in the training data (Hastie et al., 2004; Platt, 2000; Hansen & Sargent, 2001). However, compiling the SVM model for the current study required the longest time to program in Python. This was already mentioned in other studies (Yu et al., 2004; Auria & Moro, 2008; Fedorovici & Dragan, 2011) as a disadvantage that the SVM algorithm is the most expensive algorithm to compile computationally.

In addition to the longest computation time to compile SVM, there is another considerable limitation of SVM. Since the current study chose a “linear” kernel for compiling the SVM model, it was possible to implement the model to conduct practical analyses such as feature importance and predictive probability calculations at an individual level of data. However, as the data gets more complex and abundant, the SVM model may likely need to deal with non-parametric data which needs to apply non-parametric related kernels such as RBF. So far, there is no methodology to conduct feature importance and predictive probabilities for individual levels of the data when such non-parametric SVMs are applied. Therefore, it may derive limitations on using an SVM model pragmatically when a non-parametric kernel is applied.

Artificial Neural Network (ANN) Algorithm Implementation

As mentioned in the results chapter, the ANN model was compiled with 50 input nodes, 21 hidden nodes with one hidden layer, and two output nodes. The ANN model was the second-best algorithm in addition to SVM to compile college enrollment prediction modeling. The results showed that the ANN had high accuracy (0.893), specificity (0.918), and ROC-AUC (0.849) rates relative to sensitivity (0.765), precision (0.647), F_1 score (0.701), and PR-AUC (0.716). This may be reflected as having similar results to LR, DT, and SVM. However, it turns out that the ANN model is classifying imbalanced data fairly well, which has the second highest sensitivity (0.765), precision (0.647), F_1 score (0.701), and PR-AUC (0.716) values than LR and DT. In other words, the ANN model shows a better classification prediction performance for enrolled students as enrolled than LR and DT.

Like SVM, the ANN algorithm is highly prone to handle complex, high-dimensional, and imbalanced data. Similar to SVM, which requires determining the cost parameter value, choosing optional numbers of the hidden layer(s) and node(s) is necessary for ANN before applying it to the dataset. This implies that the ANN model is more flexible in adjusting to the data relative to LR and DT. As long as an optimal number of the hidden layer(s) and node(s) are applied, it is possible to anticipate that the model will perform an efficient prediction on classifying binary outcomes. It also required more computational time to compile the modeling than LR and DT but less than SVM in the python environment.

SVM and ANN Application on Test Data

This section discusses the study's primary findings, mainly focused on the two best machine learning models. As mentioned earlier, SVM and ANN were the two machine learning models that showed the best classification performance of college enrollment prediction. Both have significantly higher sensitivity, precision, F_1 score and AUC values than LR and DT models. This implied that both models are performing well in classifying and predicting students who enrolled as enrolled. Although the current study chose both data as the best models, there is enough space to think about determining the best of best model between SVM vs. ANN. Lux et al. (2016) compared SVM vs. ANN for college enrollment prediction and chose ANN as a better model, only based on the values derived from the confusion matrix. In other words, the study did not describe why ANN was better than SVM from a methodological viewpoint, not reflecting the advantages and disadvantages of the algorithm to one’s data. In fact, there are many variations and controversies on which machine learning model is better for compiling college enrollment modeling. Moreover, it is more focused on comparing the models based on the accuracies, specificity, sensitivity, precision, and ROC-AUC without considering data cross-validation and its imbalance.

Since the data are structured differently across institutions, I approached this with a broader theme regarding classification performance using highly imbalanced data. Some studies compared the classification performance between SVM vs. ANN with respect to imbalanced data classification. Ren (2012) compared the SVM vs. ANN performance based on his empirical study, which used imbalanced data from mammogram imaging. He concluded that ANN performs better on imbalanced data than SVM, showing higher accuracy, sensitivity, and precision, and ROC-AUC rates. However, other empirical studies conducted by Morares et al. (2013), Arora et al. (2010), and Ustuner (2016) stated that SVM performs better than ANN. Morares et al.(2013) and Arora et al. (2010) conducted an empirical study regarding text classification using imbalanced data and compared the performances of SVM vs. ANN. Both studies concluded that SVM is better at showing more stability with less noise on train and test data. In addition, Ustuner et al. (2016) conducted an empirical study regarding rapid-eye imaging imbalanced data and compared its performances for ANN and SVM. The study concluded that SVM was a robust, consistent, and effective classifier for imbalanced data relative to ANN.

Overall, SVM is often considered better than ANN for classification prediction on highly imbalanced data because it has a built-in mechanism for handling class imbalance, known as class weighting (Tang et al., 2002). Class weighting adjusts the relative importance of the different classes in the training process to mitigate the impact of the imbalanced distribution (Tang et al., 2002). In highly imbalanced data, the minority class typically has fewer samples than the majority class, which can lead to a biased model that predicts the majority class more frequently. This can be problematic in classification tasks where the minority class is of particular interest, such as college enrollment decisions, fraud detection, or disease diagnosis. In contrast, linear SVM with class weighting assigns higher weights to the minority class samples during training, effectively making their contribution to the optimization problem more significant. This can help to balance the influence of the different classes and improve the models' ability to identify the minority class correctly. On the other hand, ANN may struggle with highly imbalanced data if not appropriately addressed during training ANN relies on minimizing the overall prediction error, which may lead to overfitting o the majority class, causing poor performance on the minority class. It can be challenging to balance the contribution of different classes in ANN without additional techniques, such as adjusting the loss function.

Generally, SVM is a robust and effective ML algorithm for classification prediction on highly imbalanced data due to its ability to handle class imbalance with class weighting. However, determining which machine learning model is better for classification performance on imbalanced data is still controversial because other combining techniques, such as data resampling and cost-sensitive learning, can also be used to improve the performance of ML algorithm models on imbalanced data. Moreover, college enrollment data is highly imbalanced in general, and the type of data structure differs across institutions. Hence constant empirical studies for compiling predictive college enrollment modeling using various machine learning algorithms are necessary to increase the number of quantitative case studies.

Feature Importance and Predictive Probability Analyses based on SVM and ANN

The context that differs from the previous studies is that this study presented how machine learning prediction models can be used pragmatically. All past studies presented their best machine-learning models for predicting college enrollment based on the confusion matrix, derived values, and ROC-AUC. However, they did not provide any additional follow-up methods to implement how these machine learning algorithms models can be applied for practical use. Hence, I presented how to formulate the prediction probabilities using the best-selected machine learning models (i.e., SVM and ANN) and provided the probability of each student’s likelihood to enroll. This delivers significant insights to higher education administration stakeholders for setting up efficient enrollment management strategies that are highly correlated with its financial budgeting plans for its fiscal year.

Data Imbalance

It is common that college admission data involves two highly imbalanced groups (i.e., enrolled vs. non-enrolled). In general, the subset size of non-enrolled students is bigger than those who enrolled. Since the sizes of the two groups are highly imbalanced, it is important to consider how to implement these imbalanced data into the model to reduce the over/underfitting and its biases. However, most research conducted predictive modeling without considering such data imbalance traits (Antons & Maltz, 2006; Chang, 2006; Vialardi et al., 2011; Regab et al., 2014; Lux et al., 2016; Cirelli et al., 2018; Slim et al., 2019; Basu et al., 2019) and just used plain k-fold cross-validation method. The disadvantage of the plain k-fold cross-validation method is that the train data is randomly sampled k times without considering the imbalance between two groups (i.e., enrolled vs. non-enrolled). Therefore, it is highly likely that each sample has a different ratio of enrolled vs. non-enrolled records, resulting inaccurate classification prediction performance of the models. On the other hand, stratified k-fold cross-validation is sampling the original dataset k times and each sample has a similar data distribution to the original. This implies that the models can be trained and validated using each fold and be sure that the data distribution involving two groups (enrolled vs. non-enrolled) stays consistent. Hence, one primary context that differentiates the current study from the past studies is the stratified k-fold cross-validation implementation.

Limitations

For this section, I discuss the limitation of the study in terms of generalization, influential factors applied in the model, the timeframe of collected data, and the approach of using machine learning algorithms and their implementation. As mentioned in the methodology chapter, the data were collected from a 4-year non-profit private university in Midwest urban area. Hence the outcomes and findings of the current study may not be generalized to other institutions. According to Fung and Adams (2017), there are differences in students' perceptions toward types of colleges (i.e., public vs. private, 2-year vs. 4-year), locations (i.e., rural vs. urban) and this supports the claim that the current study's scope of inference is limited.

The current study reflected the ten influential factors on students' college enrollment decisions based on Perna's (2006) college choice theoretical framework. Ten factors involved students' demographics (i.e., gender, race, socioeconomic status, academic performance), high school/community context (i.e., parent education level, high school academic climate), higher education context (i.e., college proximity, institutional financial aid), and social, economic, and policy context (i.e., national unemployment and inflation rates). Since study outcomes and findings are generated based on ten factors, identified predictive models are limited to apply to other datasets involving other factors such as student engagement and major of interest, etc.

In addition to other factors influencing students' college enrollment decisions, the current study has another limitation related to the data collection timeframe. Since the data cover the cohort years from 2013 to 2019, these data reflect the events before the Covid-19 pandemic. Hence, the college enrollment prediction models generated based on the pre-pandemic period data have limitations to apply to the new data for 2020 and after. Since the test-optional policy was implemented widely across institutions during the pandemic, most institutions started to put less weight on the applicant's standardized test scores (Fair Test, 2022). In addition, the number of students who submitted their standardized test scores decreased significantly. Hence new admission policy (i.e., test-optional policy) must be considered and reflected in the model for compiling new college enrollment prediction model(s).

Another limitation of the study is related to the approach of using machine learning algorithms and their implementation. Although machine learning algorithms are prone to deal with big and high-dimensional data with less conservative implementing data, they require a long period for training and validation. Moreover, to minimize the over/underfitting, more data is the better strategy to make the model less prone to errors. Therefore, the machine learning models are generally time-expensive methodologies to generalize. However, they are easy to implement and expect highly reliable predictions once the compiling and validation stages are finalized.

Implications

Currently, we are in an environment where huge amounts of data are explosively produced in a short period, and the power of big data is growing. Hence, accurate analysis of big data that can establish a clear competitive strategy is more than anything else. Many companies use big data to analyze people's behavior and trends to develop their business strategies. And these actions are directly related to the company's performance and economic feasibility. From this point of view, the influence of decision-making by data is extensive and vital. Moreover, the higher education sector is also in the inevitable stage of dealing with big data, especially enrollment management. As mentioned in the introduction chapter, the high school graduation rate has decreased and students are applying to multiple colleges (Western Interstate Commission of Higher Education, 2020; Campbell et al., 2007). This resulted in competition growing among colleges for enrollment, and thus, institutions need to anticipate uncertainties related to budgets expected from student enrollment. In addition, the type of considerable factors influencing students' behavior related to enrollment decisions become diverse and led institutional data to grow abundant. In other words, institutions have to face managing and analyzing big data. Therefore, implementing machine learning algorithms and generating classification models for college enrollment prediction becomes crucial.

The current study identified SVM and ANN as the best models to implement for college enrollment prediction by training, validation, and testing seven years of cohort data. Moreover, the study conducted feature analysis based on these two models and identified the most important factors that impact students' enrollment decisions; Distance from Campus (i.e., proximity), High School GPA, Expected Family Contribution, Unemployment Rate, and Institutional Financial Aid Total amount. Future studies and adapt and implement these findings to support the admission, marketing, and financial aid department in setting up strategies for student recruitment. For example, student recruitment strategies can be set up in various ways to target students living closer to campus to those living further. These strategies can be combined with other strategies related to students' high school GPAs, their expected family contribution, and institutional financial aid, such as providing more discount rates (i.e., a high amount of institutional financial aid) to students who are living further with low expected family contribution but with outstanding high school GPA.

In addition to conducting feature importance analysis, the predictive probability was formulated using SVM and ANN models. This methodology provided the probability of students' likelihood to enroll based on the 46 one-hot encoded factors. Since SVM and ANN models are identified to be the best prediction models for classifying enrolled and not enrolled students, they can be applied to the incoming students' data and determine who is likely to enroll and who is not. Based on these predictive probability calculations, an institution can individually apply marketing strategies to students based on 50 factors. These can also effectively improve the diversity of institutions' student bodies by targeting students in minor communities with small populations.

Future Study

Despite limitations, the current study can be expanded to some future studies. As aforementioned, there are considerable and additional factors, such as student engagement. Student engagement is an essential factor in predicting students' college enrollment decisions (Fraysier et al., 2020). Such engagement is defined as any form of interaction that prospective students create between higher education institutions (Cole et al., 2009). This includes campus visits, attending college fairs, meeting with the admission counselors, and submitting a request for information form, etc. Hence past studies have shown that students with a significant amount of engagement and making certain types of engagement are highly likely to enroll (Fraysier et al., 2020; Peruta & Shields, 2018; Kowalik, 2011). However, this data field is immensely treated as large categorical data. Hence, conducting appropriate data-preprocessing steps is crucial before implementing it into the model. Once the data-preprocessing for student engagement data is completed, it can be implemented into compiling machine learning algorithm models for predicting students' college enrollment decisions.

In addition to the 'student engagement' factor, it is necessary to implement the 'test-optional policy' factor, which became significant after the Covid-19 pandemic. Since an institution in this study implemented the test-optional policy starting from Fall 2021 cohort, this factor must be reflected to compile the new form of college enrollment prediction model. However, since the test-optional policy-related data has been collected for only three years, it might not be sufficient to train the model(s) and apply it to test data as an upcoming new dataset. However, the new predictive college enrollment model(s) is inevitable to compile, including test-optional related factors, no later than two years from now.

Along with implementing additional factors to the prediction model(s), interactions among independent variables (i.e., influential factors on college enrollment decisions) need to be considered. Interaction refers to the effect of the combination of two or more predictor variables on the outcome variable, where the effect of one predictor variable on the outcome depends on the level of the other predictor variable(s) (Zhang, 2016). In other words, the effect of one variable on the outcome is not constant across all levels of the other variables. This is sometimes referred to as a synergistic effect, where the combined effect of two variables is greater (or less) than the sum of their individual effects (Zhang, 2016). Since the current study did not consider interactions, compiling interactions across predictor variables will be necessary for developing better classification prediction model(s). Based on the literature review chapter, there is domain knowledge in the current study regarding influential factors and their relationships to students' college enrollment decisions. In other words, prior knowledge and theory suggest that certain independent variables may interact with each other to influence the dependent variable (Zhang, 2016). Therefore, it could be necessary to analyze interactions for the compiled prediction model as an in-depth analysis. Because, in some cases, analyzing interactions may help to capture non-linear relationships and incorporate domain knowledge as the number of factors get abundant and various.

Aside from considering interactions across variables, there are additional methodologies to apply in the future analysis regarding multicollinearity assessment. While this study utilized the variance inflation factor (VIF) to detect multicollinearity, VIF measures the degree of multicollinearity between a predictor variable and all other predictor variables but does not capture the partial correlations between individual predictor variables (Kim, 2019). Therefore, VIF may not detect cases where two or more predictor variables are highly correlated with each other but not with the response variable. Nevertheless, VIF is still a useful tool for detecting multicollinearity, and it is often used in conjunction with other methods to check for multicollinearity. Hence, supplementary methods such as examining regression coefficients and computing eigenvalues of the correlation matrix can also be employed to check for multicollinearity.

Lastly, it is predictable that the student's enrollment decision behaviors and factors will get more complex and varied. Hence, this is predictable that linear modeling may not be sufficient to apply. Therefore, studies on implementing advanced model(s) are needed. I recommend applying the SVM model with the RBF kernel. This model is highly adjustable to linear and polynomial data structures in high-dimensional settings. However, the SVM model using the RBF kernel requires determining the optimal gamma parameter value and a cost parameter. The gamma parameter is a value that tunes the equation. Similar to cost parameter determination, there is no rule of thumb to decide the optimal gamma value. Therefore, testing multiple values of gamma and cost parameters as pairs is required, which takes longer to compile the final SVM model. Thus, compiling an SVM model with an RBF kernel is computationally very expensive than compiling a linear SVM model. However, it can be one of the models that give highly reliable predictions of classification performance by adding new variables (e.g., test-optional policy, engagements) to the model.

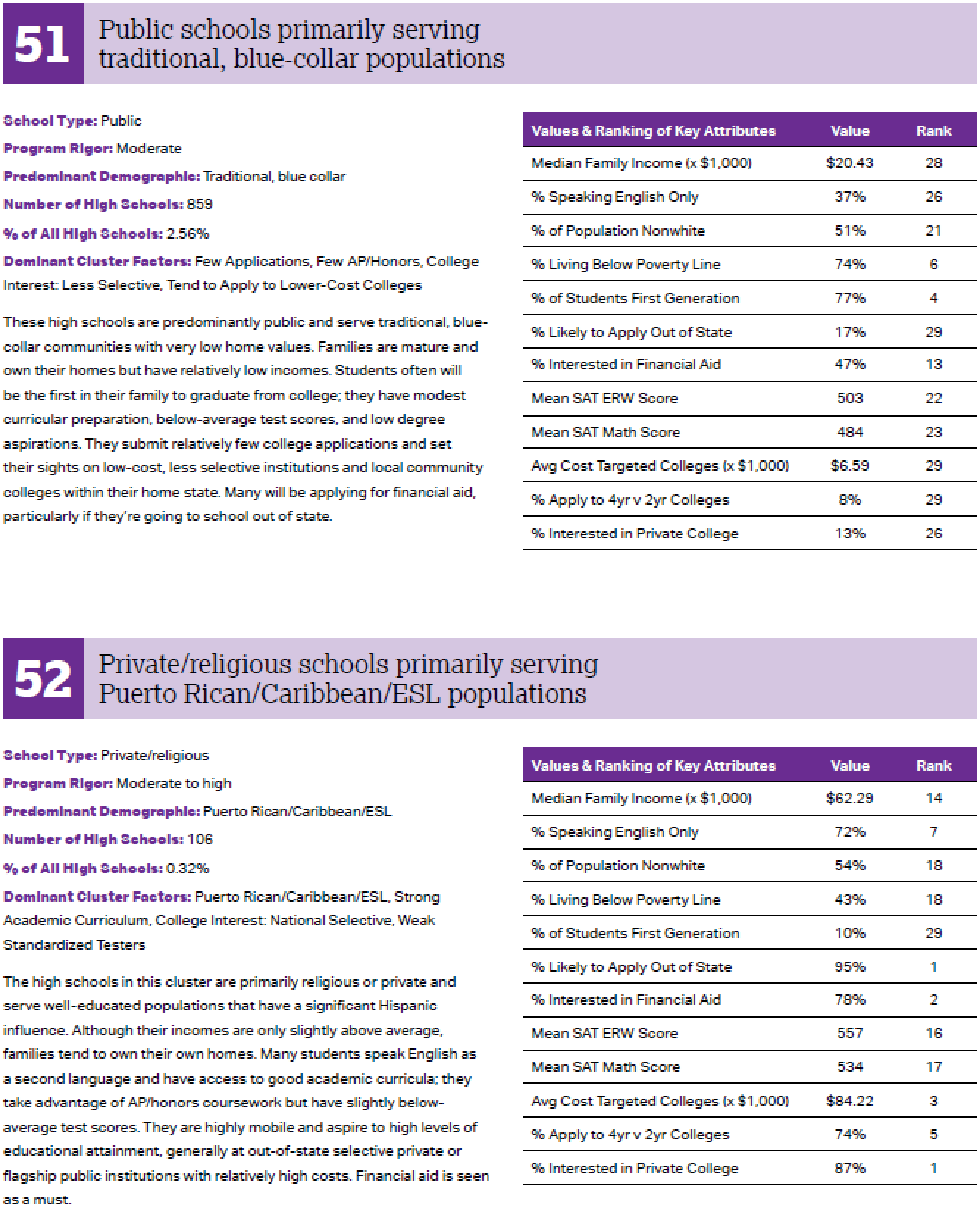
APPENDIX A

DATA VARIABLES

Variable	Meaure Type	Definition
Sex_F	Binary	Dummary Code, 1=Yes, 0=No
Sex_M	Binary	Dummary Code, 1=Yes, 0=No
Race_Asian	Binary	Dummary Code, 1=Yes, 0=No
Race_Black_AA	Binary	Dummary Code, 1=Yes, 0=No
Race_Hispanic	Binary	Dummary Code, 1=Yes, 0=No
Race_Other	Binary	Dummary Code, 1=Yes, 0=No
Race_Unknown	Binary	Dummary Code, 1=Yes, 0=No
Race_White	Binary	Dummary Code, 1=Yes, 0=No
PEL_Associates_Degree	Binary	Dummary Code, 1=Yes, 0=No
PEL_Bachelors_Degree	Binary	Dummary Code, 1=Yes, 0=No
PEL_Graduate_Degree	Binary	Dummary Code, 1=Yes, 0=No
PEL_High_School_Graduate	Binary	Dummary Code, 1=Yes, 0=No
PEL_Some_College	Binary	Dummary Code, 1=Yes, 0=No
PEL_Some_High_School	Binary	Dummary Code, 1=Yes, 0=No
SAT_ACT_SuperScore	Conintuous	Higest standardized test score, Range from 1 to 36
HSGPA	Conintuous	Raw High School GPA Scores
Inst_Fin_Aid_Total	Conintuous	Total Institutional Financial Aid Amount
Fed_Efc	Conintuous	Federal Expected Family Contribution
Distance_from_Campus	Conintuous	Proximity
Unemployment_Rate	Conintuous	National Unemployment Rate by month, Year
Inflation_Rate	Conintuous	Inflation Rate by Month, Year
Enroll_Decision	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_51	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_52	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_53	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_54	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_55	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_56	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_57	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_58	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_59	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_60	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_61	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_62	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_63	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_64	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_65	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_66	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_67	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_68	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_69	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_70	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_71	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_72	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_73	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_74	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_75	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_76	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_77	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_78	Binary	Dummary Code, 1=Yes, 0=No
HSCluster_79	Binary	Dummary Code, 1=Yes, 0=No

APPENDIX B

DESCRIPTIONS OF HIGH SCHOOL ACADEMIC CLIMATE CLUSTERS



51 Public schools primarily serving traditional, blue-collar populations

School Type: Public

Program Rigor: Moderate

Predominant Demographic: Traditional, blue collar

Number of High Schools: 859

% of All High Schools: 2.56%

Dominant Cluster Factors: Few Applications, Few AP/Honors, College Interest: Less Selective, Tend to Apply to Lower-Cost Colleges

These high schools are predominantly public and serve traditional, blue-collar communities with very low home values. Families are mature and own their homes but have relatively low incomes. Students often will be the first in their family to graduate from college; they have modest curricular preparation, below-average test scores, and low degree aspirations. They submit relatively few college applications and set their sights on low-cost, less selective institutions and local community colleges within their home state. Many will be applying for financial aid, particularly if they're going to school out of state.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$20.43	28
% Speaking English Only	37%	26
% of Population Nonwhite	51%	21
% Living Below Poverty Line	74%	6
% of Students First Generation	77%	4
% Likely to Apply Out of State	17%	29
% Interested in Financial Aid	47%	13
Mean SAT ERW Score	503	22
Mean SAT Math Score	484	23
Avg Cost Targeted Colleges (x $1,000)	$6.59	29
% Apply to 4yr v 2yr Colleges	8%	29
% Interested in Private College	13%	26

52 Private/religious schools primarily serving Puerto Rican/Caribbean/ESL populations

School Type: Private/religious

Program Rigor: Moderate to high

Predominant Demographic: Puerto Rican/Caribbean/ESL

Number of High Schools: 106

% of All High Schools: 0.32%

Dominant Cluster Factors: Puerto Rican/Caribbean/ESL, Strong Academic Curriculum, College Interest: National Selective, Weak Standardized Testers

The high schools in this cluster are primarily religious or private and serve well-educated populations that have a significant Hispanic influence. Although their incomes are only slightly above average, families tend to own their own homes. Many students speak English as a second language and have access to good academic curricula; they take advantage of AP/honors coursework but have slightly below-average test scores. They are highly mobile and aspire to high levels of educational attainment, generally at out-of-state selective private or flagship public institutions with relatively high costs. Financial aid is seen as a must.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$62.29	14
% Speaking English Only	72%	7
% of Population Nonwhite	54%	18
% Living Below Poverty Line	43%	18
% of Students First Generation	10%	29
% Likely to Apply Out of State	95%	1
% Interested in Financial Aid	78%	2
Mean SAT ERW Score	557	16
Mean SAT Math Score	534	17
Avg Cost Targeted Colleges (x $1,000)	$84.22	3
% Apply to 4yr v 2yr Colleges	74%	5
% Interested in Private College	87%	1

53 Comprehensive public/religious schools primarily serving traditional, blue-collar communities

School Type: Private/religious

Program Rigor: Low to moderate

Predominant Demographic: Traditional, blue collar

Number of High Schools: 1,011

% of All High Schools: 3.01%

Dominant Cluster Factors: Religious Curriculum, Few AP/Honors, College Interest: Less Selective, Lower Ability

These high schools are often religiously affiliated and serve middle-class communities with a mix of professional, managerial, and blue-collar households. Most families are acquainted with college, but only a modest proportion of households includes a college graduate. Students tend to get good grades, but their test scores are below average, and their involvement in AP and honors courses is minimal. Their degree aspirations are quite low, and their college choices tend toward less selective and lower-cost, church-related institutions close to home. Many will be applying for financial aid.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$45.08	20
% Speaking English Only	47%	23
% of Population Nonwhite	50%	23
% Living Below Poverty Line	55%	11
% of Students First Generation	52%	12
% Likely to Apply Out of State	39%	22
% Interested in Financial Aid	33%	17
Mean SAT ERW Score	525	20
Mean SAT Math Score	502	21
Avg Cost Targeted Colleges (x $1,000)	$23.83	27
% Apply to 4yr v 2yr Colleges	28%	28
% Interested in Private College	24%	24

54 Public schools in rural settings primarily serving African American and Hispanic populations

School Type: Public, rural

Program Rigor: Low

Predominant Demographic: African American and Hispanic

Number of High Schools: 433

% of All High Schools: 1.29%

Dominant Cluster Factors: Primarily African American, Black Inner City, First Generation College-Going, Not Athletic Participant

These high schools serve predominantly rural, working-class African American and Hispanic families at the lowest end of the economic scale. Few parents have any experience with college. Students have access to a general curriculum with few AP or honors opportunities; their test scores are at or near the bottom. Although they are willing to look out of state and apply to moderately selective institutions as well as local two-year and technical colleges, students from these schools seem to have low aspirations and little guidance or information regarding financial aid.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$16.00	29
% Speaking English Only	64%	14
% of Population Nonwhite	93%	1
% Living Below Poverty Line	91%	1
% of Students First Generation	77%	5
% Likely to Apply Out of State	44%	20
% Interested in Financial Aid	11%	23
Mean SAT ERW Score	439	29
Mean SAT Math Score	428	29
Avg Cost Targeted Colleges (x $1,000)	$38.84	22
% Apply to 4yr v 2yr Colleges	37%	22
% Interested in Private College	6%	29

55 Private/religious schools predominantly serving males from racially diverse populations

School Type: Private/religious

Program Rigor: Low to moderate

Predominant Demographic: Racially diverse males

Number of High Schools: 949

% of All High Schools: 2.83%

Dominant Cluster Factors: College Prep School, Affluent with Modest Aspirations, College Interest: Lower-Cost Public, Relatively Low Grades

The high schools in this cluster are primarily private or religiously affiliated and serve predominantly male, racially mixed populations from homes with modest, above-average incomes. Most parents have attended college and hold predominantly professional or managerial positions. Although education is a community value, student participation in AP and honors courses, standardized test scores, and aspirations beyond high school are all below average. Students are willing to consider going to college out of state, and they tend to apply to moderately priced and relatively selective institutions.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$70.11	7
% Speaking English Only	59%	19
% of Population Nonwhite	56%	17
% Living Below Poverty Line	39%	20
% of Students First Generation	36%	17
% Likely to Apply Out of State	64%	11
% Interested in Financial Aid	20%	22
Mean SAT ERW Score	537	18
Mean SAT Math Score	531	18
Avg Cost Targeted Colleges (x $1,000)	$59.61	15
% Apply to 4yr v 2yr Colleges	53%	13
% Interested in Private College	50%	16

56 Public/private schools serving racially diverse populations with a strong interest in athletics

School Type: Public/private

Program Rigor: Low to moderate

Predominant Demographic: Racially diverse, athletic interest

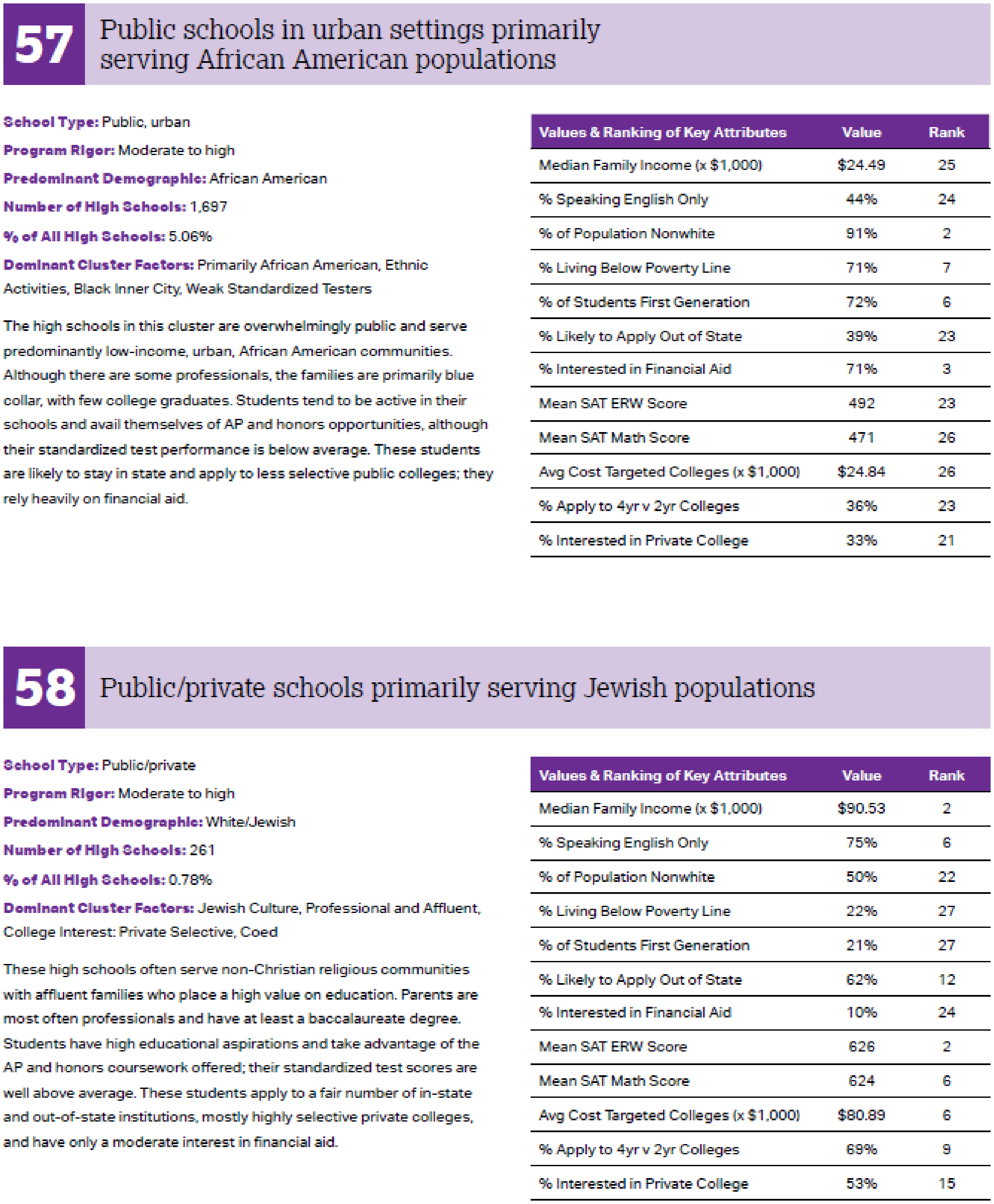
Number of High Schools: 969

% of All High Schools: 2.89%

Dominant Cluster Factors: Athletic Achievements, Affluent and Mobile, Few, Highly Targeted Applications , Few AP/Honors

These high schools, sometimes religious, serve solidly middle-class, moderately diverse, and slightly older communities with a mix of professional, managerial, and blue-collar households; the students may have strong athletic traditions. Most families have a parent with at least some college experience. Although students aren't involved in many AP or honors courses, they have access to strong STEM courses and perform at an above-average level on standardized tests. They don't apply to many institutions and tend to prefer selective private institutions with higher costs, often outside their home state. Interest in financial aid is moderate.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$62.21	15
% Speaking English Only	62%	16
% of Population Nonwhite	63%	11
% Living Below Poverty Line	39%	21
% of Students First Generation	37%	16
% Likely to Apply Out of State	85%	3
% Interested in Financial Aid	31%	18
Mean SAT ERW Score	565	12
Mean SAT Math Score	577	9
Avg Cost Targeted Colleges (x $1,000)	$74.31	9
% Apply to 4yr v 2yr Colleges	69%	10
% Interested in Private College	65%	10



57 Public schools in urban settings primarily serving African American populations

School Type: Public, urban

Program Rigor: Moderate to high

Predominant Demographic: African American

Number of High Schools: 1,697

% of All High Schools: 5.06%

Dominant Cluster Factors: Primarily African American, Ethnic Activities, Black Inner City, Weak Standardized Testers

The high schools in this cluster are overwhelmingly public and serve predominantly low-income, urban, African American communities. Although there are some professionals, the families are primarily blue collar, with few college graduates. Students tend to be active in their schools and avail themselves of AP and honors opportunities, although their standardized test performance is below average. These students are likely to stay in state and apply to less selective public colleges; they rely heavily on financial aid.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$24.49	25
% Speaking English Only	44%	24
% of Population Nonwhite	91%	2
% Living Below Poverty Line	71%	7
% of Students First Generation	72%	6
% Likely to Apply Out of State	39%	23
% Interested in Financial Aid	71%	3
Mean SAT ERW Score	492	23
Mean SAT Math Score	471	26
Avg Cost Targeted Colleges (x $1,000)	$24.84	26
% Apply to 4yr v 2yr Colleges	36%	23
% Interested in Private College	33%	21

58 Public/private schools primarily serving Jewish populations

School Type: Public/private

Program Rigor: Moderate to high

Predominant Demographic: White/Jewish

Number of High Schools: 261

% of All High Schools: 0.78%

Dominant Cluster Factors: Jewish Culture, Professional and Affluent, College Interest: Private Selective, Coed

These high schools often serve non-Christian religious communities with affluent families who place a high value on education. Parents are most often professionals and have at least a baccalaureate degree. Students have high educational aspirations and take advantage of the AP and honors coursework offered; their standardized test scores are well above average. These students apply to a fair number of in-state and out-of-state institutions, mostly highly selective private colleges, and have only a moderate interest in financial aid.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$90.53	2
% Speaking English Only	75%	6
% of Population Nonwhite	50%	22
% Living Below Poverty Line	22%	27
% of Students First Generation	21%	27
% Likely to Apply Out of State	62%	12
% Interested in Financial Aid	10%	24
Mean SAT ERW Score	626	2
Mean SAT Math Score	624	6
Avg Cost Targeted Colleges (x $1,000)	$80.89	6
% Apply to 4yr v 2yr Colleges	69%	9
% Interested in Private College	53%	15

59 Public schools in suburban settings primarily serving white, blue-collar populations

School Type: Public, suburban

Program Rigor: Moderate to high

Predominant Demographic: White, blue collar

Number of High Schools: 5,405

% of All High Schools: 16.11%

Dominant Cluster Factors: High Grades Relative to Test Scores, Working Class, College Interest: Low-Cost Publics, In-State Colleges

These are suburban public high schools serving older, economically depressed, white, blue-collar communities. Most parents have only a high school education and are employed in skilled trades or other nonprofessional occupations. Students tend to perform well in the classroom, take modest advantage of the advanced courses offered, and have very modest educational aspirations and standardized test scores. These students don't apply to many institutions and tend to favor less selective public institutions and community colleges in their home state. Financial aid will be a large factor.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$30.02	24
% Speaking English Only	25%	29
% of Population Nonwhite	30%	28
% Living Below Poverty Line	58%	10
% of Students First Generation	68%	7
% Likely to Apply Out of State	18%	28
% Interested in Financial Aid	64%	5
Mean SAT ERW Score	540	17
Mean SAT Math Score	526	19
Avg Cost Targeted Colleges (x $1,000)	$26.38	25
% Apply to 4yr v 2yr Colleges	30%	27
% Interested in Private College	30%	22

60 Private schools primarily serving Jewish female populations

School Type: Private

Program Rigor: Low to moderate

Predominant Demographic: White, predominately Jewish females

Number of High Schools: 159

% of All High Schools: 0.47%

Dominant Cluster Factors: Jewish Culture, Traditional Single Gender, Affluent, Interested in Religious Education

The high schools in this cluster are primarily private or sectarian. They serve mostly women with professional, college-educated parents who are often from non-Christian communities. Household incomes and home values are above average. Students are academically oriented and perform well in class and on standardized tests, although they are generally uninvolved in AP and honors coursework. They tend to focus their applications on a few moderately priced, relatively selective private institutions and to stay close to home and have a below-average interest in financial aid.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$71.51	6
% Speaking English Only	82%	4
% of Population Nonwhite	62%	12
% Living Below Poverty Line	54%	14
% of Students First Generation	24%	25
% Likely to Apply Out of State	21%	27
% Interested in Financial Aid	10%	25
Mean SAT ERW Score	602	8
Mean SAT Math Score	566	11
Avg Cost Targeted Colleges (x $1,000)	$56.41	16
% Apply to 4yr v 2yr Colleges	52%	15
% Interested in Private College	11%	27

61 Private schools in urban settings serving racially diverse populations

School Type: Private, urban

Program Rigor: Low

Predominant Demographic: Racially diverse

Number of High Schools: 303

% of All High Schools: 0.90%

Dominant Cluster Factors: Strong Academic Curriculum, Small Private, Large Asian Population, First Generation College-Going

These high schools are predominantly private and serve older, racially mixed, inner-city communities where some of the population speaks English as a second language. There is an almost equal split between professional, managerial, and blue-collar occupations. Students are exposed to college prep curricula but not AP and honors courses. Standardized test scores are below average, with the lowest scores on language-related sections. Students aspire to earn graduate degrees and apply to a small number of moderately selective private institutions. They seem disinterested in financial aid despite very average family incomes, which may suggest that families place a very high priority on higher education.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$56.60	16
% Speaking English Only	70%	9
% of Population Nonwhite	57%	14
% Living Below Poverty Line	54%	13
% of Students First Generation	25%	24
% Likely to Apply Out of State	45%	18
% Interested in Financial Aid	2%	29
Mean SAT ERW Score	468	27
Mean SAT Math Score	546	15
Avg Cost Targeted Colleges (x $1,000)	$65.70	12
% Apply to 4yr v 2yr Colleges	53%	14
% Interested in Private College	18%	25

62 Public schools serving Hispanic populations with traditional values

School Type: Public

Program Rigor: High

Predominant Demographic: Hispanic, traditional

Number of High Schools: 433

% of All High Schools: 1.29%

Dominant Cluster Factors: Hispanic, Diverse Low Income, Traditional Catholic Culture, High Educational Aspirations

The high schools in this cluster serve predominantly lower-middle-class, bilingual Hispanic families with strong traditional values. Many parents have had some experience in higher education, which is reflected in their mix of professional, managerial, and blue-collar occupations. Students take a range of college prep classes and frequently have access to AP and honors-level courses, but their standardized test scores are below average. These students are moderately mobile and tend to apply to lower-cost, relatively selective private institutions where financial aid will be important.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$46.66	18
% Speaking English Only	97%	1
% of Population Nonwhite	44%	25
% Living Below Poverty Line	62%	9
% of Students First Generation	60%	8
% Likely to Apply Out of State	55%	15
% Interested in Financial Aid	53%	11
Mean SAT ERW Score	537	19
Mean SAT Math Score	519	20
Avg Cost Targeted Colleges (x $1,000)	$45.03	19
% Apply to 4yr v 2yr Colleges	60%	12
% Interested in Private College	57%	12

63 Public schools in urban settings primarily serving Hispanic/ESL and African American populations

School Type: Private, urban

Program Rigor: Moderate

Predominant Demographic: Hispanic/ESL and African American

Number of High Schools: 809

% of All High Schools: 2.41%

Dominant Cluster Factors: Hispanic, African American, First Generation College-Going, Relatively Low Grades

These public high schools serve an inner-city mix of established nonwhite populations, about half of whom speak English as a second language. Families often include younger children; and parents generally have below-average incomes, don't own their homes, have completed high school or some college, and are in blue-collar or lower-level professional jobs. Students have moderate educational goals and are involved in some AP and honors coursework, but they score consistently below average on admission tests. They tend to look at in-state public colleges or reasonably priced and moderately selective private institutions from which they will expect financial aid.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$32.81	22
% Speaking English Only	90%	3
% of Population Nonwhite	87%	4
% Living Below Poverty Line	78%	5
% of Students First Generation	86%	2
% Likely to Apply Out of State	36%	25
% Interested in Financial Aid	54%	10
Mean SAT ERW Score	481	26
Mean SAT Math Score	478	25
Avg Cost Targeted Colleges (x $1,000)	$45.53	18
% Apply to 4yr v 2yr Colleges	36%	25
% Interested in Private College	33%	20

64 Public schools primarily serving Asian/ESL populations

School Type: Public

Program Rigor: Moderate to high

Predominant Demographic: Asian/ESL

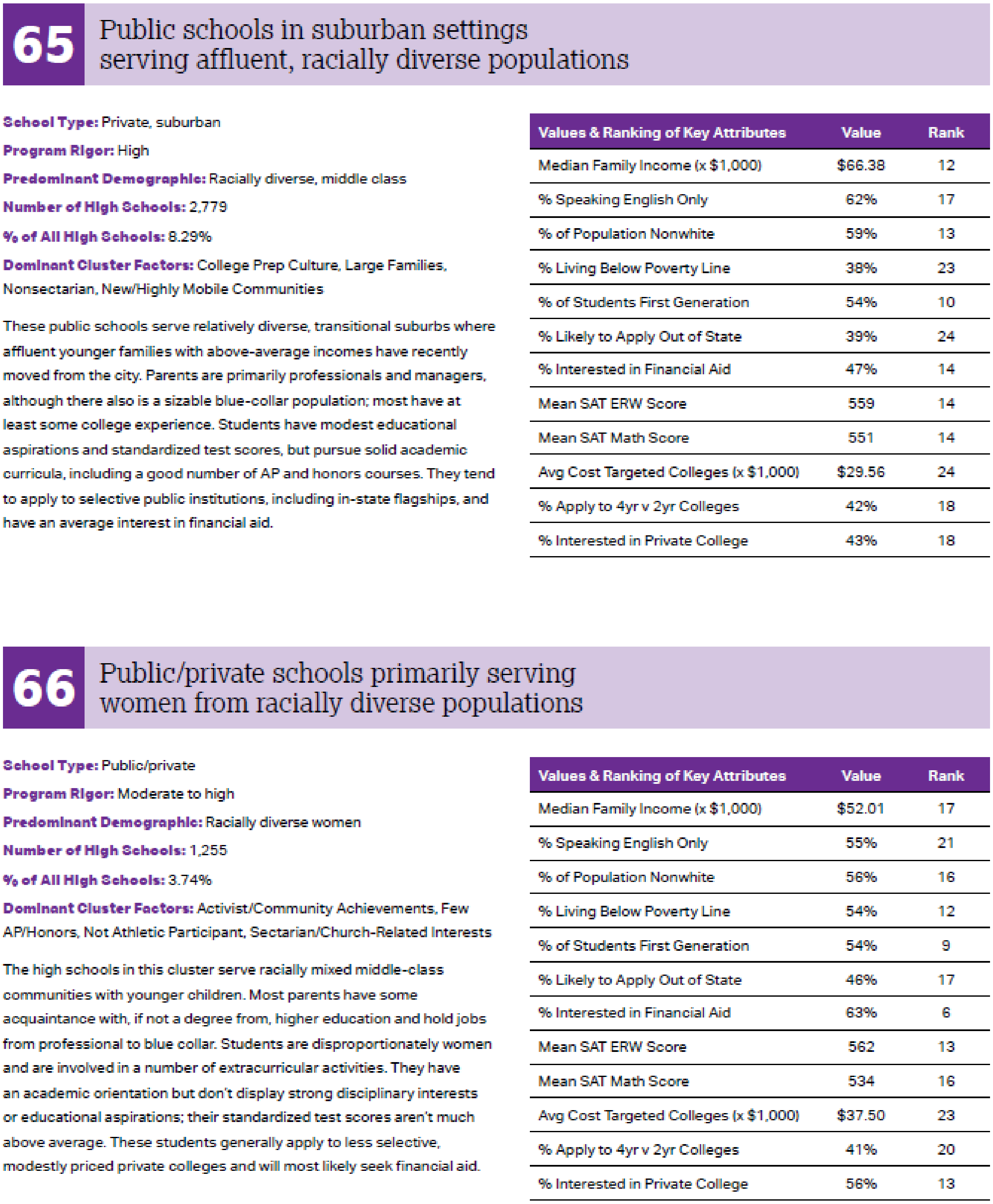
Number of High Schools: 638

% of All High Schools: 1.90%

Dominant Cluster Factors: Large Asian/ESL population, College Prep School, College Interest: Selective Cost Conscious, Nonsectarian

The high schools in this cluster are mostly public and serve predominantly younger, Asian families, many of whom are bilingual. The parents have broad experience with higher education and hold professional or managerial positions that provide well-above-average incomes. Students pursue both math/science and liberal arts curricula, take full advantage of AP and honors courses, and score well on standardized tests. Although these students aren't overly mobile and have only an average interest in financial aid, they do consider cost when looking at higher education options and will likely apply to many different colleges across a range of selectivity.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$73.97	5
% Speaking English Only	81%	5
% of Population Nonwhite	77%	5
% Living Below Poverty Line	40%	19
% of Students First Generation	44%	13
% Likely to Apply Out of State	64%	10
% Interested in Financial Aid	26%	19
Mean SAT ERW Score	617	4
Mean SAT Math Score	675	2
Avg Cost Targeted Colleges (x $1,000)	$74.11	10
% Apply to 4yr v 2yr Colleges	75%	4
% Interested in Private College	68%	9



65 Public schools in suburban settings serving affluent, racially diverse populations

School Type: Private, suburban

Program Rigor: High

Predominant Demographic: Racially diverse, middle class

Number of High Schools: 2,779

% of All High Schools: 8.29%

Dominant Cluster Factors: College Prep Culture, Large Families, Nonsectarian, New/Highly Mobile Communities

These public schools serve relatively diverse, transitional suburbs where affluent younger families with above-average incomes have recently moved from the city. Parents are primarily professionals and managers, although there also is a sizable blue-collar population; most have at least some college experience. Students have modest educational aspirations and standardized test scores, but pursue solid academic curricula, including a good number of AP and honors courses. They tend to apply to selective public institutions, including in-state flagships, and have an average interest in financial aid.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$66.38	12
% Speaking English Only	62%	17
% of Population Nonwhite	59%	13
% Living Below Poverty Line	38%	23
% of Students First Generation	54%	10
% Likely to Apply Out of State	39%	24
% Interested in Financial Aid	47%	14
Mean SAT ERW Score	559	14
Mean SAT Math Score	551	14
Avg Cost Targeted Colleges (x $1,000)	$29.56	24
% Apply to 4yr v 2yr Colleges	42%	18
% Interested in Private College	43%	18

66 Public/private schools primarily serving women from racially diverse populations

School Type: Public/private

Program Rigor: Moderate to high

Predominant Demographic: Racially diverse women

Number of High Schools: 1,255

% of All High Schools: 3.74%

Dominant Cluster Factors: Activist/Community Achievements, Few AP/Honors, Not Athletic Participant, Sectarian/Church-Related Interests

The high schools in this cluster serve racially mixed middle-class communities with younger children. Most parents have some acquaintance with, if not a degree from, higher education and hold jobs from professional to blue collar. Students are disproportionately women and are involved in a number of extracurricular activities. They have an academic orientation but don't display strong disciplinary interests or educational aspirations; their standardized test scores aren't much above average. These students generally apply to less selective, modestly priced private colleges and will most likely seek financial aid.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$52.01	17
% Speaking English Only	55%	21
% of Population Nonwhite	56%	16
% Living Below Poverty Line	54%	12
% of Students First Generation	54%	9
% Likely to Apply Out of State	46%	17
% Interested in Financial Aid	63%	6
Mean SAT ERW Score	562	13
Mean SAT Math Score	534	16
Avg Cost Targeted Colleges (x $1,000)	$37.50	23
% Apply to 4yr v 2yr Colleges	41%	20
% Interested in Private College	56%	13

67 Religious/private schools primarily serving women, upper-middle-class populations

School Type: Private

Program Rigor: High

Predominant Demographic: Upper-middle-class women

Number of High Schools: 1,558

% of All High Schools: 4.65%

Dominant Cluster Factors: Strong Academics, College Interest: National Selective, Leadership/Organizational Achievements, Artistic Interests

The schools in this cluster are most often religiously affiliated and predominantly serve women from older, upper-middle-class communities. Most parents have at least some college experience and are either professionals or managers. Students are academically oriented and involved in a number of activities; they take solid curricula in both math/science and AP/honors; and they score above average on standardized tests. These students have fairly high educational aspirations, are relatively mobile, and apply to a good number of selective private colleges; many will seek financial aid.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$67.06	9
% Speaking English Only	65%	13
% of Population Nonwhite	66%	9
% Living Below Poverty Line	45%	17
% of Students First Generation	41%	15
% Likely to Apply Out of State	67%	9
% Interested in Financial Aid	61%	7
Mean SAT ERW Score	590	10
Mean SAT Math Score	560	12
Avg Cost Targeted Colleges (x $1,000)	$77.55	7
% Apply to 4yr v 2yr Colleges	69%	8
% Interested in Private College	81%	4

68 Religious schools primarily serving Catholic populations

School Type: Private/religious

Program Rigor: High

Predominant Demographic: Catholic

Number of High Schools: 854

% of All High Schools: 2.55%

Dominant Cluster Factors: Catholic Culture, Selective Cost Conscious, Highly Educated, Coed

These high schools are almost exclusively religious and predominantly Catholic. They serve communities with extensive home ownership and household incomes well above average. Almost all parents have some college experience, and most are either professionals or managers. Students are active in their communities and athletics; they tend to have moderate educational aspirations, solid involvement in AP and honors coursework, and above-average test scores. These students apply to a fair number of schools, mostly selective, moderately priced private and sectarian colleges in their home state. Financial aid is on the minds of a majority.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$77.66	4
% Speaking English Only	58%	20
% of Population Nonwhite	52%	20
% Living Below Poverty Line	27%	26
% of Students First Generation	29%	20
% Likely to Apply Out of State	60%	14
% Interested in Financial Aid	38%	16
Mean SAT ERW Score	596	9
Mean SAT Math Score	586	8
Avg Cost Targeted Colleges (x $1,000)	$70.90	11
% Apply to 4yr v 2yr Colleges	67%	11
% Interested in Private College	70%	7

69 Public schools primarily serving African American populations

School Type: Public/private

Program Rigor: Low

Predominant Demographic: Inner city/African American

Number of High Schools: 705

% of All High Schools: 2.1%

Dominant Cluster Factors: Primarily African American, First-Generation College-Going, Black Inner City, Less Academically Prepared

These high schools serve very low-income, predominantly African American communities. Although the largest proportion of parents hold blue-collar jobs and have only a high school education, there's also a noticeable professional and managerial presence. Students tend to be active in school and have an academic orientation, although participation in advanced coursework is quite low, and test scores are near the bottom. Some students will look out of state at somewhat selective, moderately priced private institutions, but many will choose a public two- or four-year college close to home. Financial aid will be essential for most.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$21.51	27
% Speaking English Only	63%	15
% of Population Nonwhite	90%	3
% Living Below Poverty Line	83%	2
% of Students First Generation	82%	3
% Likely to Apply Out of State	45%	19
% Interested in Financial Aid	52%	12
Mean SAT ERW Score	456	28
Mean SAT Math Score	443	28
Avg Cost Targeted Colleges (x $1,000)	$44.01	21
% Apply to 4yr v 2yr Colleges	33%	26
% Interested in Private College	24%	23

70 Public schools primarily serving affluent suburban populations

School Type: Public, suburban

Program Rigor: High

Predominant Demographic: Affluent professionals

Number of High Schools: 1,291

% of All High Schools: 3.85%

Dominant Cluster Factors: Professional and Affluent, Good Standardized Testers, Activist/Community Achievements, National Selective

These primarily public schools serve established, very affluent suburban communities. Parents overwhelmingly are in professional or managerial positions, with over half holding a postbaccalaureate degree. Students have access to strong curricula, take advantage of AP and honors coursework, are active and involved in a variety of activities, and perform very well on standardized tests. Overwhelmingly committed to earning a degree, they apply to many highly selective public and private colleges, both in and out of state. Despite the costs associated with their college choices, slightly less than half of these students will seek financial aid.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$90.94	1
% Speaking English Only	62%	18
% of Population Nonwhite	52%	19
% Living Below Poverty Line	19%	29
% of Students First Generation	22%	26
% Likely to Apply Out of State	77%	8
% Interested in Financial Aid	21%	21
Mean SAT ERW Score	632	1
Mean SAT Math Score	638	4
Avg Cost Targeted Colleges (x $1,000)	$81.89	5
% Apply to 4yr v 2yr Colleges	73%	6
% Interested in Private College	80%	5

71 Public/private/religious schools primarily serving Puerto Rican/Caribbean/ESL populations

School Type: Private/religious

Program Rigor: Low

Predominant Demographic: Puerto Rican/Caribbean/ESL

Number of High Schools: 187

% of All High Schools: 0.56%

Dominant Cluster Factors: Puerto Rican/Caribbean/ESL, Focused/Early Decision, College Interest: Flagship Public, Few AP/Honors

The high schools in this cluster, about one-third of which are private or religiously affiliated, serve low-income Hispanic communities with large families. Although the largest proportion are in blue-collar occupations, most parents have had some college. Students tend toward less challenging coursework but perform well in their classes; a few get involved with AP and honors courses. Their standardized test scores are near the bottom. These students tend to be rather focused on their college choices, often looking at either public flagships or somewhat selective, moderately priced private institutions where financial aid would be a must.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$31.40	23
% Speaking English Only	68%	11
% of Population Nonwhite	68%	7
% Living Below Poverty Line	67%	8
% of Students First Generation	35%	18
% Likely to Apply Out of State	77%	7
% Interested in Financial Aid	81%	1
Mean SAT ERW Score	485	25
Mean SAT Math Score	454	27
Avg Cost Targeted Colleges (x $1,000)	$44.90	20
% Apply to 4yr v 2yr Colleges	36%	24
% Interested in Private College	68%	8

72 Homeschoolers and private/religious schools primarily serving upper-middle-class Christian populations

School Type: Private/religious

Program Rigor: Moderate

Predominant Demographic: Christian, upper middle class

Number of High Schools: 2,221

% of All High Schools: 6.62%

Dominant Cluster Factors: Religious Activities, Christian Culture, College Interest: Sectarian, Relatively High Grades

These schools, which are predominantly Christian affiliated and may include homeschoolers, serve upper-middle-class communities where most families own their homes. Parents work in a variety of vocations, and almost all have at least some experience with higher education. Students generally are exposed to good, sometimes above-average curricula, are involved in AP and honors coursework, and attain above-average standardized test scores. Their educational aspirations are very modest; these students apply to fewer schools than most and generally consider less selective, private, church-related institutions. Their interest in financial aid is about average.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$66.82	11
% Speaking English Only	50%	22
% of Population Nonwhite	49%	24
% Living Below Poverty Line	36%	24
% of Students First Generation	26%	21
% Likely to Apply Out of State	48%	16
% Interested in Financial Aid	39%	15
Mean SAT ERW Score	584	11
Mean SAT Math Score	568	10
Avg Cost Targeted Colleges (x $1,000)	$54.12	17
% Apply to 4yr v 2yr Colleges	47%	16
% Interested in Private College	46%	17

73 Public schools in urban settings primarily serving Hispanic (particularly Mexican) populations

School Type: Public, urban

Program Rigor: Moderate to high

Predominant Demographic: Hispanic, particularly Mexican

Number of High Schools: 1,212

% of All High Schools: 3.61%

Dominant Cluster Factors: Mexican, Large Families, Primarily First-Generation College-Going, Diverse Low Income

The schools in this cluster are generally public and serve urban families with modest incomes and large families. Although there's some diversity, families are largely blue collar, with large Mexican and other Hispanic populations, speak English as a second language, and have little or no experience with college. Although the students test below average, they avail themselves of academic opportunities and frequently seek out AP and honors coursework. They apply to a reasonable number of public two- and four-year colleges, mostly within their home state, as well as some less selective and relatively low-cost private, primarily Catholic, institutions. Financial aid is seen as essential to attending college.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$28.33	26
% Speaking English Only	94%	2
% of Population Nonwhite	81%	6
% Living Below Poverty Line	81%	4
% of Students First Generation	92%	1
% Likely to Apply Out of State	32%	26
% Interested in Financial Aid	65%	4
Mean SAT ERW Score	490	24
Mean SAT Math Score	483	24
Avg Cost Targeted Colleges (x $1,000)	$23.71	28
% Apply to 4yr v 2yr Colleges	41%	19
% Interested in Private College	35%	19

74 Private schools primarily serving Asian/ESL populations

School Type: Private

Program Rigor: High

Predominant Demographic: Asian/ESL

Number of High Schools: 886

% of All High Schools: 2.64%

Dominant Cluster Factors: Large Asian/ESL population, College Interest: Private Selective, Higher Ability, Leadership/Organizational Achievements

These schools are most often private and serve highly educated, relatively small, middle-class families. These families are more likely to be professional than blue collar, and the largest ethnic group is Asian. Students seek out strong curricula, although their involvement in AP and honors courses is modest. They have extremely high educational aspirations and score at or near the top on standardized tests. These students are highly mobile and apply to a number of institutions, generally to some of the most selective and expensive private colleges. Despite only modest income levels, their interest in financial aid is slightly below average.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	66.86	10
% Speaking English Only	70%	10
% of Population Nonwhite	64%	10
% Living Below Poverty Line	45%	16
% of Students First Generation	25%	23
% Likely to Apply Out of State	90%	2
% Interested in Financial Aid	24%	20
Mean SAT ERW Score	610	5
Mean SAT Math Score	660	3
Avg Cost Targeted Colleges (x $1,000)	$92.27	1
% Apply to 4yr v 2yr Colleges	78%	2
% Interested in Private College	84%	2

75 Public schools in rural settings primarily serving middle-class populations with traditional values

School Type: Public, rural

Program Rigor: Moderate to high

Predominant Demographic: Middle class, traditional values

Number of High Schools: 1,967

% of All High Schools: 5.86%

Dominant Cluster Factors: Tend to Stay Close to Home, Athletic Achievements, College Interest: Sectarian/Catholic, Low Aspirations

The schools in this cluster are overwhelmingly public and represent well-established small town and rural communities where almost everyone owns a home and has a comfortable income. Most parents have traditional values and some experience with college; they are also vocationally diverse. Students generally take basic college prep curricula and only modestly get involved in AP and honors coursework. Their educational aspirations are low, and they generally receive average test scores. These students are drawn to moderately priced colleges that are close to home and somewhat selective, where financial aid will be available.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$68.56	8
% Speaking English Only	41%	25
% of Population Nonwhite	31%	27
% Living Below Poverty Line	22%	28
% of Students First Generation	53%	11
% Likely to Apply Out of State	43%	21
% Interested in Financial Aid	56%	8
Mean SAT ERW Score	559	15
Mean SAT Math Score	555	13
Avg Cost Targeted Colleges (x $1,000)	$63.41	13
% Apply to 4yr v 2yr Colleges	45%	17
% Interested in Private College	54%	14

76 Private schools primarily serving affluent, racially diverse populations

School Type: Private

Program Rigor: Moderate

Predominant Demographic: Affluent and diverse private high schools

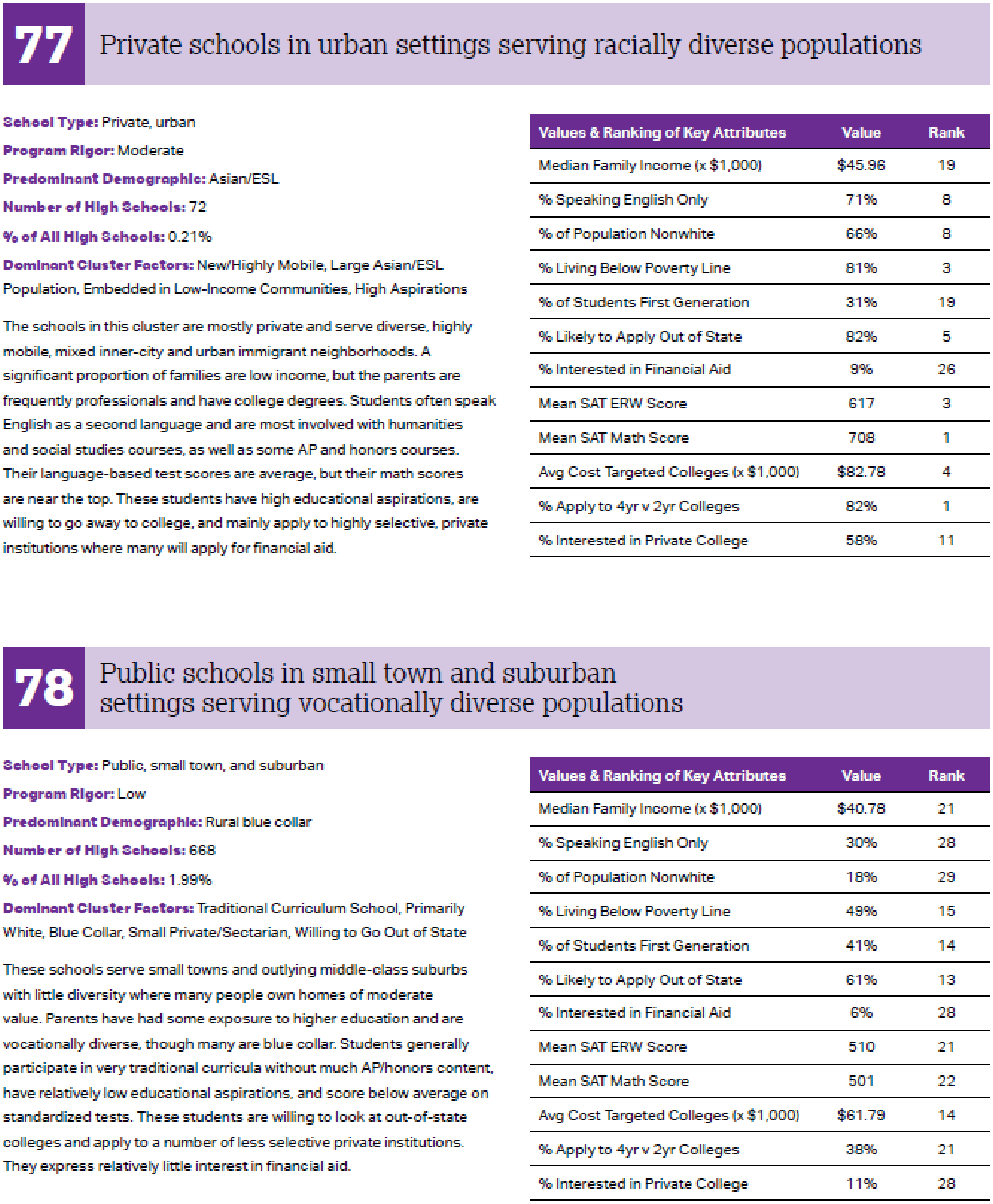
Number of High Schools: 476

% of All High Schools: 1.42%

Dominant Cluster Factors: College Prep School/Liberal Arts, High Educational Aspirations, College Interest: Small, Selective Privates, Affluent Diversity

Overwhelmingly private, the schools in this cluster serve somewhat racially/ethnically mixed, upper-income families with few children. Parents are almost all professionals or managers and are highly educated. Students have good curricula, which include solid math and science and some AP and honors courses. They score at or near the top on standardized tests and generally aspire to postbaccalaureate education. They are willing to travel, and they consider a large number of colleges—generally highly selective and expensive private institutions where only some will apply for financial aid.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$82.33	3
% Speaking English Only	66%	12
% of Population Nonwhite	57%	15
% Living Below Poverty Line	38%	22
% of Students First Generation	21%	28
% Likely to Apply Out of State	84%	4
% Interested in Financial Aid	8%	27
Mean SAT ERW Score	606	7
Mean SAT Math Score	629	5
Avg Cost Targeted Colleges (x $1,000)	$90.84	2
% Apply to 4yr v 2yr Colleges	77%	3
% Interested in Private College	73%	6



77 Private schools in urban settings serving racially diverse populations

School Type: Private, urban

Program Rigor: Moderate

Predominant Demographic: Asian/ESL

Number of High Schools: 72

% of All High Schools: 0.21%

Dominant Cluster Factors: New/Highly Mobile, Large Asian/ESL Population, Embedded in Low-Income Communities, High Aspirations

The schools in this cluster are mostly private and serve diverse, highly mobile, mixed inner-city and urban immigrant neighborhoods. A significant proportion of families are low income, but the parents are frequently professionals and have college degrees. Students often speak English as a second language and are most involved with humanities and social studies courses, as well as some AP and honors courses. Their language-based test scores are average, but their math scores are near the top. These students have high educational aspirations, are willing to go away to college, and mainly apply to highly selective, private institutions where many will apply for financial aid.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$45.96	19
% Speaking English Only	71%	8
% of Population Nonwhite	66%	8
% Living Below Poverty Line	81%	3
% of Students First Generation	31%	19
% Likely to Apply Out of State	82%	5
% Interested in Financial Aid	9%	26
Mean SAT ERW Score	617	3
Mean SAT Math Score	708	1
Avg Cost Targeted Colleges (x $1,000)	$82.78	4
% Apply to 4yr v 2yr Colleges	82%	1
% Interested in Private College	58%	11

78 Public schools in small town and suburban settings serving vocationally diverse populations

School Type: Public, small town, and suburban

Program Rigor: Low

Predominant Demographic: Rural blue collar

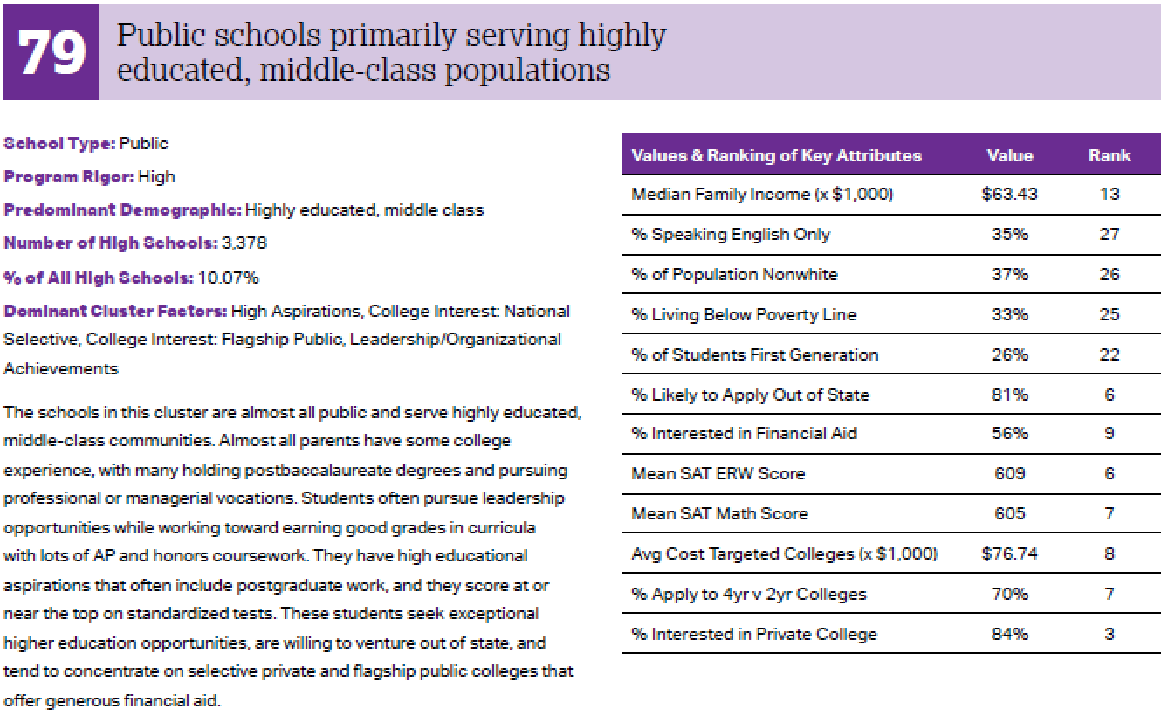
Number of High Schools: 668

% of All High Schools: 1.99%

Dominant Cluster Factors: Traditional Curriculum School, Primarily White, Blue Collar, Small Private/Sectarian, Willing to Go Out of State

These schools serve small towns and outlying middle-class suburbs with little diversity where many people own homes of moderate value. Parents have had some exposure to higher education and are vocationally diverse, though many are blue collar. Students generally participate in very traditional curricula without much AP/honors content, have relatively low educational aspirations, and score below average on standardized tests. These students are willing to look at out-of-state colleges and apply to a number of less selective private institutions. They express relatively little interest in financial aid.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$40.78	21
% Speaking English Only	30%	28
% of Population Nonwhite	18%	29
% Living Below Poverty Line	49%	15
% of Students First Generation	41%	14
% Likely to Apply Out of State	61%	13
% Interested in Financial Aid	6%	28
Mean SAT ERW Score	510	21
Mean SAT Math Score	501	22
Avg Cost Targeted Colleges (x $1,000)	$61.79	14
% Apply to 4yr v 2yr Colleges	38%	21
% Interested in Private College	11%	28



79 Public schools primarily serving highly educated, middle-class populations

School Type: Public

Program Rigor: High

Predominant Demographic: Highly educated, middle class

Number of High Schools: 3,378

% of All High Schools: 10.07%

Dominant Cluster Factors: High Aspirations, College Interest: National Selective, College Interest: Flagship Public, Leadership/Organizational Achievements

The schools in this cluster are almost all public and serve highly educated, middle-class communities. Almost all parents have some college experience, with many holding postbaccalaureate degrees and pursuing professional or managerial vocations. Students often pursue leadership opportunities while working toward earning good grades in curricula with lots of AP and honors coursework. They have high educational aspirations that often include postgraduate work, and they score at or near the top on standardized tests. These students seek exceptional higher education opportunities, are willing to venture out of state, and tend to concentrate on selective private and flagship public colleges that offer generous financial aid.

Values & Ranking of Key Attributes	Value	Rank
Median Family Income (x $1,000)	$63.43	13
% Speaking English Only	35%	27
% of Population Nonwhite	37%	26
% Living Below Poverty Line	33%	25
% of Students First Generation	26%	22
% Likely to Apply Out of State	81%	6
% Interested in Financial Aid	56%	9
Mean SAT ERW Score	609	6
Mean SAT Math Score	605	7
Avg Cost Targeted Colleges (x $1,000)	$76.74	8
% Apply to 4yr v 2yr Colleges	70%	7
% Interested in Private College	84%	3

APPENDIX C

PYTHON CODE OPEN SOURCE

```
import pandas as pd
import numpy as np
import seaborn as sns
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.model_selection import train_test_split
from sklearn.model_selection import StratifiedKFold
from matplotlib import pyplot as plt
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import confusion_matrix, roc_auc_score ,roc_curve, auc
from sklearn.metrics import precision_recall_curve
from sklearn.linear_model import LogisticRegression
from sklearn import tree
from sklearn import svm
from sklearn.neural_network import MLPClassifier

###Train/Validation Dataset Application###
#Train/Validation Dataset Launch#
data = pd.read_csv("/Users/Final Data/TrainValidation.csv")
X = data.drop(['Term', 'STRM', 'Race_White', 'Sex_Female',
'PEL_Bachelors_Degree','HSCluster_79','Enroll_Decision'], axis=1)
y = data['Enroll_Decision']

#Multicollinearity Assessment using VIF#
vif_data = pd.DataFrame()
vif_data["feature"] = X.columns
vif_data["VIF"] = [variance_inflation_factor(X.values, i)
                   for i in range(len(X.columns))]
print(vif_data)

# Compile LR model with Stratified 10-Fold Cross-Validation for Test/Validation Data#
kf = StratifiedKFold(n_splits=10, shuffle=True)
#due to the train_test_split function, validation dataset is treated as a test data#
pred_test_full =0
cv_score =[]
i=1
for train_index,test_index in kf.split(X,y):
   print('{} of KFold {}'.format(i,kf.n_splits))
   xtr,xvl = X.loc[train_index],X.loc[test_index]
   ytr,yvl = y.loc[train_index],y.loc[test_index]
lr = LogisticRegression()
lr.fit(xtr,ytr)
```

```
#testing ROC AUC values to see whether stratified 10-fold CV works properly#
score = roc_auc_score(yvl,lr.predict(xvl))
print('ROC AUC score:',score)
cv_score.append(score)
pred_test = lr.predict_proba(x_test)[:,1]
pred_test_full +=pred_test
i+=1

#Concatenated Confusion Matrix of LR#
print('Concatenated Confusion matrix\n',confusion_matrix(yvl,lr.predict(xvl)))
LRConcatCM= confusion_matrix(yvl,lr.predict(xvl))
sns.heatmap(LRConcatCM, annot=True, fmt=".0F", cmap=colormap)
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.title(‘Concatenated Confusion Matrix for Logistic Regression’)
plt.show()

#Mean ROC-AUC for LR#
proba = lr.predict_proba(xvl)[:,1]
fpr,tpr, threshold = roc_curve(yvl,proba)
roc_auc_ = auc(fpr,tpr)
plt.figure(figsize=(12, 7))
plt.title(‘Mean ROC Curve for Logistic Regression', size=20)
plt.plot(fpr, tpr, ‘r’, label = 'AUC (LR)= %0.3f' % roc_auc_)
plt.xlabel('False Positive Rate', size=14)
plt.ylabel('True Positive Rate', size=14)
plt.legend();

# Mean PR-AUC for LR#
precision_lr, recall_lr, _ = precision_recall_curve(yvl,proba)
auc_lr = auc(recall_lr, precision_lr)
plt.figure(figsize=(12, 7))
plt.plot(recall_lr, precision_lr, label=f'AUC (LR) = {auc_lr:.3f}',color='red')
plt.title('Mean Precision-Recall Curve for Logistic Regression’, size=20)
plt.xlabel('Recall', size=14)
plt.ylabel('Precision', size=14)
plt.legend();
plt.show()

# Compile DT model with Stratified 10-Fold Cross-Validation for Test/Validation Data#
kf = StratifiedKFold(n_splits=10, shuffle=True)
pred_test_full =0
cv_score =[]
i=1
```

```
for train_index,test_index in kf.split(X,y):
    print('{} of KFold {}'.format(i,kf.n_splits))
    xtr,xvl = X.loc[train_index],X.loc[test_index]
    ytr,yvl = y.loc[train_index],y.loc[test_index]

     dt = tree.DecisionTreeClassifier()
     dt.fit(xtr,ytr)
     score = roc_auc_score(yvl,lr.predict(xvl))
     print('ROC AUC score:',score)
     cv_score.append(score)
     pred_test = dt.predict_proba(x_test)[:,1]
     pred_test_full +=pred_test
     i+=1

#Concatenated Confusion Matrix of LR#
DTConcatCM= confusion_matrix(yvl,dt.predict(xvl))
sns.heatmap(DTConcatCM, annot=True, fmt=".0F", cmap=colormap)
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.title(‘Concatenated Confusion Matrix for Decision Tree’)
plt.show()

#ROC-AUC for DT#
proba = dt.predict_proba(xvl)[:,1]
fpr,tpr, threshold = roc_curve(yvl,proba)
roc_auc_ = auc(fpr,tpr)
plt.figure(figsize=(12, 7))
plt.title('Mean ROC Curve for Decision Tree', size=20)
plt.plot(fpr, tpr, ‘r’, label = 'AUC (DT)= %0.3f' % roc_auc_)
plt.xlabel('False Positive Rate', size=14)
plt.ylabel('True Positive Rate', size=14)
plt.legend();

#For DT Model – Compile Mean PR-AUC#
precision_dt, recall_dt, _ = precision_recall_curve(yvl,proba)
auc_dt = auc(recall_dt, precision_dt)
plt.figure(figsize=(12, 7))
plt.plot(recall_dt, precision_dt, label=f'AUC (DT) = {auc_dt:.3f}',color='red')
plt.title('Mean Precision-Recall Curve for Decision Tree', size=20)
plt.xlabel('Recall', size=14)
plt.ylabel('Precision', size=14)
plt.legend();
plt.show()
```

```
# Compile SVM model with Stratified 10-Fold Cross-Validation for Test/Validation Data#
kf = StratifiedKFold(n_splits=10, shuffle=True)
pred_test_full =0
cv_score =[]
i=1
for train_index,test_index in kf.split(X,y):
    print('{} of KFold {}'.format(i,kf.n_splits))
    xtr,xvl = X.loc[train_index],X.loc[test_index]
    ytr,yvl = y.loc[train_index],y.loc[test_index]

     svm = svm.SVC(kernel=’linear’, C=10.0)

svm.fit(xtr,ytr)
score = roc_auc_score(yvl,lr.predict(xvl))
print('ROC AUC score:',score)
cv_score.append(score)
pred_test = svm.predict_proba(x_test)[:,1]
pred_test_full +=pred_test
i+=1

#Concatenated Confusion Matrix of SVM#
SVMConcatCM= confusion_matrix(yvl,svm.predict(xvl))
sns.heatmap(SVMConcatCM, annot=True, fmt=".0F",cmap=colormap)
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.title('Concatenated Confusion Matrix for Support Vector Machine')
plt.show()

#Mean ROC-AUC for SVM#
proba = svm.predict_proba(xvl)[:,1]
fpr,tpr, threshold = roc_curve(yvl,proba)
roc_auc_ = auc(fpr,tpr)
plt.figure(figsize=(12, 7))
plt.title('Mean ROC Curve for Support Vector Machine', size=20)
plt.plot(fpr, tpr, ‘r’, label = 'AUC (SVM)= %0.3f' % roc_auc_)
plt.xlabel('False Positive Rate', size=14)
plt.ylabel('True Positive Rate', size=14)
plt.legend();
plt.show()

#For SVM Model – Compile Mean PR-AUC#
precision_svm, recall_svm, _ = precision_recall_curve(yvl,proba)
auc_svm = auc(recall_svm, precision_svm)
plt.figure(figsize=(12, 7))
plt.plot(recall_svm, precision_svm, label=f'AUC (SVM) = {auc_svm:.3f}',color='red')
plt.title('Mean Precision-Recall Curve for Support Vector Machine', size=20)
plt.xlabel('Recall', size=14)
plt.ylabel('Precision', size=14)
plt.legend();
plt.show()

# Compile ANN model with Stratified 10-Fold Cross-Validation for Test/Validation Data#
kf = StratifiedKFold(n_splits=10, shuffle=True)
pred_test_full =0
cv_score =[]
i=1
for train_index,test_index in kf.split(X,y):
   print('{} of KFold {}'.format(i,kf.n_splits))
   xtr,xvl = X.loc[train_index],X.loc[test_index]
   ytr,yvl = y.loc[train_index],y.loc[test_index]

    ann = MLPClassifier(solver=’lbfgs’, hidden_layer_sizes=(27,))
ann.fit(xtr,ytr)
score = roc_auc_score(yvl,ann.predict(xvl))
print('ROC AUC score:',score)
cv_score.append(score)
pred_test = ann.predict_proba(x_test)[:,1]
pred_test_full +=pred_test
i+=1

#Concatenated Confusion Matrix of ANN#
ANNConcatCM= confusion_matrix(yvl,ann.predict(xvl))
sns.heatmap(ANNConcatCM, annot=True, fmt=".0F",cmap=colormap)
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.title(‘Concatenated Confusion Matrix for Artificial Neural Network’)
plt.show()

#Mean ROC-AUC for ANN#
proba = ann.predict_proba(xvl)[:,1]
fpr,tpr, threshold = roc_curve(yvl,proba)
roc_auc_ = auc(fpr,tpr)
plt.figure(figsize=(12, 7))
plt.title(‘Mean ROC Curve for Artificial Neural Network’, size=20)
plt.plot(fpr, tpr, ‘r’, label = 'AUC (ANN)= %0.3f' % roc_auc_)
plt.xlabel('False Positive Rate', size=14)
plt.ylabel('True Positive Rate', size=14)
plt.legend();
plt.show()

#For ANN Model – Compile Mean PR-AUC#
precision_ann, recall_ann, _ = precision_recall_curve(yvl,proba)
auc_ann = auc(recall_ann, precision_ann)
plt.figure(figsize=(12, 7))
plt.plot(recall_ann, precision_ann, label=f'AUC (ANN) = {auc_lr:.3f}',color='red')
plt.title('Mean Precision-Recall Curve for Artificial Neural Network’, size=20)
plt.xlabel('Recall', size=14)
plt.ylabel('Precision', size=14)
plt.legend();
plt.show()

###Test Dataset Application###
#Test Dataset Launch#
data = pd.read_csv("/Users/Final Data/Test.csv")
X = data.drop(['Term', 'STRM', 'Race_White', 'Sex_Female',
'PEL_Bachelors_Degree','HSCluster_79','Enroll_Decision'], axis=1)
y = data['Enroll_Decision']

#Compile Support Vector Machine (SVM) Model#
svm = svm.SVC(kernel=’linear’, C=10.0)
svm.fit(X,y)

#Confusion Matrix of SVM#
SVMCM= confusion_matrix(y,svm.predict(X))
sns.heatmap(SVMCM, annot=True,fmt=".0F",cmap=colormap)
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.title('Confusion Matrix for Support Vector Machine for Test Data')
plt.show()

#ROC-AUC for SVM#
proba = svm.predict_proba(X)[:,1]
fpr,tpr, threshold = roc_curve(y,proba)
roc_auc_ = auc(fpr,tpr)
plt.figure(figsize=(12, 7))
plt.title('ROC Curve for Support Vector Machine on Test Data', size=20)
plt.plot(fpr, tpr, ‘r’, label = 'AUC (SVM)= %0.3f' % roc_auc_)
plt.xlabel('False Positive Rate', size=14)
plt.ylabel('True Positive Rate', size=14)
plt.legend();
plt.show()
```

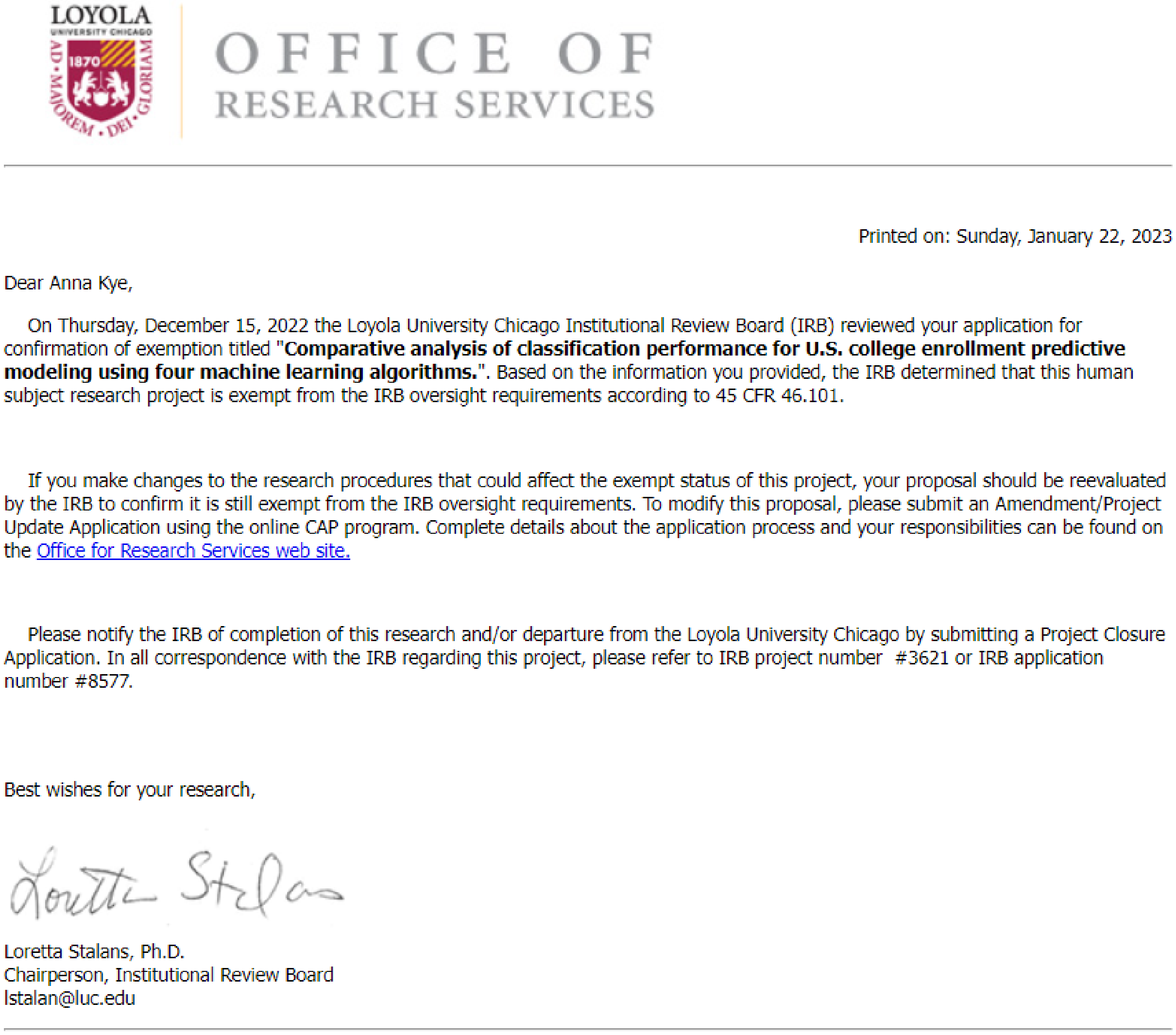
```
#For SVM Model – Compile PR-AUC#
precision_svm, recall_svm, _ = precision_recall_curve(y,proba)
auc_svm = auc(recall_svm, precision_svm)
plt.figure(figsize=(12, 7))
plt.plot(recall_svm, precision_svm, label=f'AUC (SVM) = {auc_svm:.3f}',color='red')
plt.title('Precision-Recall Curve for Support Vector Machine on Test Data', size=20)
plt.xlabel('Recall', size=14)
plt.ylabel('Precision', size=14)
plt.legend();
plt.show()

# Compile ANN Model on Test Data#
ann = MLPClassifier(solver=’lbfgs’, hidden_layer_sizes=(27,))
ann.fit(X,y)

#Concatenated Confusion Matrix of ANN#
ANNCM= confusion_matrix(y,ann.predict(X))
sns.heatmap(ANNCM, annot=True, fmt=".0F", cmap=colormap)
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.title(‘Concatenated Confusion Matrix for Artificial Neural Network’)
plt.show()

#ROC-AUC for ANN#
proba = ann.predict_proba(X)[:,1]
fpr,tpr, threshold = roc_curve(y,proba)
roc_auc_ = auc(fpr,tpr)
plt.figure(figsize=(12, 7))
plt.title('ROC Curve for Artificial Neural Network on Test Data', size=20)
plt.plot(fpr, tpr, ‘r’, label = 'AUC (ANN)= %0.3f' % roc_auc_)
plt.xlabel('False Positive Rate', size=14)
plt.ylabel('True Positive Rate', size=14)
plt.legend();
plt.show()

#For ANN Model – Compile Mean PR-AUC#
precision_ann, recall_ann, _ = precision_recall_curve(y,proba)
auc_ann = auc(recall_ann, precision_ann)
plt.figure(figsize=(12, 7))
plt.plot(recall_ann, precision_ann, label=f'AUC (ANN) = {auc_ann:.3f}',color='red')
plt.title('Precision-Recall Curve for Artificial Neural Network on Test Data’, size=20)
plt.xlabel('Recall', size=14)
plt.ylabel('Precision', size=14)
plt.legend();
plt.show()

###Feature Importance Application using SVM - Drop down column###
data = pd.read_csv("/Users/Final Data/TrainValidationTest.csv")
X = data.drop(['Term', 'STRM', 'Race_White', 'Sex_Female',
'PEL_Bachelors_Degree','HSCluster_79','Enroll_Decision'], axis=1)
y = data['Enroll_Decision']

kf = StratifiedKFold(n_splits=10, shuffle=True)
pred_valid_full =0
cv_score =[]
i=1
for train_index,test_index in kf.split(X,y):
    print('{} of KFold {}'.format(i,kf.n_splits))
    xtr,xvl = X.loc[train_index],X.loc[test_index]
    ytr,yvl = y.loc[train_index],y.loc[test_index]
    svm = svm.SVC(kernel=’linear’, C=10.0)
def dropcol_importances(svm, xtr, ytr, kf)
    svm_ = clone(svm)
    baseline = cross_val_score(svm_, xtr, ytr, scoring='accuracy', kf=kf)
    imp = []
    for col in xtr.columns:
            X = xtr.drop(col, axis=1)
            svm_ = clone(svm)
            oob = cross_val_score(svm_, xtr, ytr, scoring='accuracy', kf=kf)
            imp.append(baseline - oob)
        imp = np.array(imp)
        importance = pd.DataFrame(
                imp, index = xtr.columns)
        importnce.columns = ["kf_{}".format(i) for i in range(kf)]
drop_col_impt = dropcol_importances(svm, xtr, ytr, kf)
drop_col_importance = pd.DataFrame({'features': xtr.columns.tolist(),
                                    "drop_col_importance":
drop_col_imp.mean(axis=1).values}).sort_values('drop_col_importance', ascending=False)
drop_col_importance

###Feature Importance Application using ANN - Drop down column###
data = pd.read_csv("/Users/Final Data/TrainValidationTest.csv")
X = data.drop(['Term', 'STRM', 'Race_White', 'Sex_Female',
'PEL_Bachelors_Degree','HSCluster_79','Enroll_Decision'], axis=1)
y = data['Enroll_Decision']

kf = StratifiedKFold(n_splits=10, shuffle=True)
pred_valid_full =0
```

```
cv_score =[]
i=1
for train_index,test_index in kf.split(X,y):
    print('{} of KFold {}'.format(i,kf.n_splits))
    xtr,xvl = X.loc[train_index],X.loc[test_index]
    ytr,yvl = y.loc[train_index],y.loc[test_index]
    ann = MLPClassifier(solver=’lbfgs’, hidden_layer_sizes=(27,))
def dropcol_importances(ann, xtr, ytr, kf)
    svm_ = clone(ann)
    baseline = cross_val_score(svm_, X, ytr, scoring='accuracy', kf=kf)
    imp = []
    for col in xtr.columns:
            X = xtr.drop(col, axis=1)
            svm_ = clone(svm)
            oob = cross_val_score(ann_, X, ytr, scoring='accuracy', kf=kf)
            imp.append(baseline - oob)
        imp = np.array(imp)
        importance = pd.DataFrame(
                imp, index = xtr.columns)
        importnce.columns = ["kf_{}".format(i) for i in range(kf)]

drop_col_impt = dropcol_importances(ann, xtr, ytr, kf)
drop_col_importance = pd.DataFrame({'features': xtr.columns.tolist(),
                                    "drop_col_importance":
drop_col_imp.mean(axis=1).values}).sort_values('drop_col_importance', ascending=False)
drop_col_importance

###Predictive Probability Application###
data = pd.read_csv("/Users/Final Data/TrainValidationTest.csv")
X = data.drop(['Term', 'STRM', 'Race_White', 'Sex_Female',
'PEL_Bachelors_Degree','HSCluster_79','Enroll_Decision'], axis=1)
y = data['Enroll_Decision']

#Preditive probability compile using SVM#
svm = svm.SVC(kernel=’linear’, C=10.0)
clf=CalibratedClassifierCV(svm)
svm_clf=clf.fit(X,y)
y_proba=clf.predict_proba(X)
print(clf.predict_proba(X))
predicted=clf.predict_proba(X)
csv=pd.DataFrame(predicted, columns=['Enrolled','NotErnolled'])
csv.to_csv("/Users/predictiveprobability_svm.csv",index=False)
```

```
#Predictive probability compile using ANN#
ann = MLPClassifier(solver=’lbfgs’, hidden_layer_sizes=(27,))
ann_clf=ann.fit(X,y)
print(ann_clf.predict_proba(X))
predicted=ann_clf.predict_proba(X)
csv=pd.DataFrame(predicted, columns=['Enrolled','NotEnrolled'])
csv.to_csv("/Users/predictiveprobability_ann.csv",index=False)
```

APPENDIX D

PROOF OF IRB EXEMPT



LOYOLA UNIVERSITY CHICAGO

OFFICE OF RESEARCH SERVICES

Printed on: Sunday, January 22, 2023

Dear Anna Kye,

On Thursday, December 15, 2022 the Loyola University Chicago Institutional Review Board (IRB) reviewed your application for confirmation of exemption titled "**Comparative analysis of classification performance for U.S. college enrollment predictive modeling using four machine learning algorithms.**". Based on the information you provided, the IRB determined that this human subject research project is exempt from the IRB oversight requirements according to 45 CFR 46.101.

If you make changes to the research procedures that could affect the exempt status of this project, your proposal should be reevaluated by the IRB to confirm it is still exempt from the IRB oversight requirements. To modify this proposal, please submit an Amendment/Project Update Application using the online CAP program. Complete details about the application process and your responsibilities can be found on the Office for Research Services web site.

Please notify the IRB of completion of this research and/or departure from the Loyola University Chicago by submitting a Project Closure Application. In all correspondence with the IRB regarding this project, please refer to IRB project number #3621 or IRB application number #8577.

Best wishes for your research,

Loretta Stalans

Loretta Stalans, Ph.D.
Chairperson, Institutional Review Board
lstalan@luc.edu

BIBLIOGRAPHY

ACT. (2018). *Guide to the 2018 ACT/SAT concordance.* https://www.act.org/content/dam/act/unsecured/documents/pdfs/ACT-SAT-Concordance.pdf

Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college*. U.S. Department of Education. http://www2.ed.gov/rschstat/research/pubs/toolboxrevisit/toolbox.pdf

Allensworth, E. M., & Clark, K. (2020). High school GPAs and ACT Scores as predictors of college completion: Examining assumptions about consistency across high schools. *Educational Researcher, 49(3)*, 198–211.

Allison, P. (1999). *Multiple regression: A primer (Research methods and statistics).* Thousand Oaks, CA: Pine Forge Press.

Alpaydin, E. (2011). Machine learning. *WIREs Computational Statistics, 3*(2), 195-203.

Antons, C., & Maltz, E. N. (2006). Expanding the role of institutional research at small private universities: A case study in enrollment management using data mining. *New Directions for Institutional Research,* 2006, 69-81.

Archibald, R. B., & Feldman, D. H. (2010). *Why Does College Cost So Much?* New York: Oxford University Press.

Arisholm, E., Briand, L. C., & Johannessen, E. B. (2010). A systematic and comprehensive investigation of methods to build and evaluate fault prediction models. *The Journal of Systems and Software, 83*, 2-17.

Arora, S., Bhattacharfee, D., Nasipuri, M., Malik, L., Kundu, M., & Basu. D. K. (2010). Performance comparison of SVM and ANN for handwritten Devanagari character recognition. *International Journal of Computer Science Issues, 7*(3), 1-10.

Astin, A. W. (1965). Effect of different college environments on the vocational choices of high aptitude students. *Journal of Counseling Psychology, 12*(1), 28–34.

Astin, A.W., Astin, H.S., & Lindholm, J.A. (2011). Assessing students' spiritual and religious qualities. *Journal of College Student Development 52*(1), 39-61.

Attewell, P., & Domina, T. (2008). Raising the bar: Curricular intensity and academic performance. *Educational Evaluation and Policy Analysis, 30*(1), 51–71.

Auria, L., & Moro, R. A. (2008). Support vector machines (SVM) as a technique for solvency analysis. DIW Berlin Discussion Paper No. 811. http://dx.doi.org/10.2139/ssrn.1424949

Ayer, T., Chhatwal, J., Alagoz, O., Kahn, C. E., Jr, Woods, R. W., & Burnside, E. S. (2010). Informatics in radiology: comparison of logistic regression and artificial neural network models in breast cancer risk estimation. *Radiographics: a review publication of the Radiological Society of North America, Inc, 30*(1), 13–22.

Basu, K., Basu, T., Buckmire, R., & Lal, N. (2019). Predictive models of student college commitment decisions using machine learning. *Data, 4(2),* 65.

Batesman, M., & Spruill, D. (1996). Student decision-making: insights from the college choice process. *College Student Journal, 30,* 182-186.

Baum, S. (2001). College education: Who can afford it? In M. B. Paulsen & J. C. Smart (Eds.), *The finance of higher education: Theory, research, policy and practice* (pp. 55-95). New York, NY: Algora.

Beattie, I. R. (2002). Are all 'adolescent econometricians' created equal? Racial, class, and gender differences in college enrollment. *Sociology of Education, 75,* 19-43.

Bejou, D., Wray, B., & Ingram, T. N. (1996). Determinants of relationship quality: An artificial neural network analysis. *Journal of Business Research, 36*(2), 137-143.

Bekkar, M., Kjemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications, 3*(10), 27-38.

Belzil, C., & Hansen, J. (2003). Structural estimates of the intergenerational education correlation. *Journal of Applied Econometrics, 18*(6), 679-696.

Betts, J. R., & McFarland, L. L. (1995). Safe port in a storm: The impact of labor market conditions on community college enrollments. *The Journal of Human Resources, 30*(4), 741-765.

Bhardwa, S. (2017, June 6). Why do students go to university and how do they choose which one? timeshighereducation.com.https://www.timeshighereducation.com/student/news/why-do-students-go-universityand-how-do-they-choose-which-one#survey-answer

Bifulco, R., Fletcher, J. M., & Ross, S. L. (2011). The effect of classmate characteristics on post-secondary outcomes: evidence from the add health. *American Economic Journal: Economic Policy, 3*(1), 25–53.

Bingham, M. A., & Solverson, N. W. (2016). Using enrollment data to predict retention rate. *Journal of Student Affairs Research and Practice, 53*(1), 51-64.

Bloom, J. (2007). (Mis)reading social class in the journey towards college: Youth development in urban America. *Teachers College Record, 109*(2), 343–368.

Bourdieu, P. (1986). The forms of capital. In J. C. Richardson (Ed.), *Handbook of theory and research for the sociology of education.* New York: Greenwood Press.

Breiman, L. (2001). Random Forests. *Machine Learning 45*, 5–32.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). *Classification and regression trees*. Chapman and Hall, Wadsworth, NY: Routledge.

Brown v. Board of Education of Topeka, 347 U.S. 483 (1954).

Bruggink, T. H., & Gambhir, V. (1996). Statistical models for college admission and enrollment: A case study for a selective liberal arts college. *Research in Higher Education 37*(2), 221-240.

Bundick, B., & Pollard, E. (2019). The rise and fall of college tuition inflation. *Federal Reserve Bank of Kansas City Economic Review, 104*(1), 57-75.

Cabrera, A. F. (1994). Logistic regression analysis in higher education: An applied perspective. In John C. Smart (Ed.) *Higher Education: Handbook for the Study of Higher Education Volume 10*, pp. 225-256. New York, NY: Agathon.

Caelen, O. A. (2017). Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence, 81*, 429–450.

Cameron, S., & Heckman, J. (1998). Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of American males. *Journal of Political Economy*, *106*, 262-333.

Cameron, S. & Heckman, J. (2001). The dynamics of educational attainment for Black, Hispanic and White males. *Journal of Political Economy, 109*, 455-499.

Campbell, F., Beasley, L., Eland, J. & Rumpus, A. (2007) *Hearing the student voice: promoting and encouraging the effective use of the student voice to enhance professional development in learning, teaching and assessment within higher education*. Project Report for Escalate. http://escalate.ac.uk/downloads/3911.pdf.

Carbonaro, W., Ellison, B. J., & Covay, E. (2011). Gender inequalities in the college pipeline. *Social Science Research, 40*(1), 120-135.

Cavart, C. L., & Khoshgoftaar, T. M. (2019). Threshold based optimization of performance metrics with severely imbalanced big security data [Paper presentation]. 31st IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR (pp. 1328-1334).

Cerda, P., Varoquaux, G. & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Mach Learning, 107*, 1477–1494.

Chang, L. (2006). Applying data mining to predict college admissions yield: A case study. *New Directions for Institutional Research, 2006*(131), 2006.

Chapman, D.W. (1981) A model of student college choice. *Journal of Higher Education, 52,* 490-505.

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data, 7*, 52.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), 2 Jan. 2020, p. NA. *Gale OneFile: Health and Medicine,* link.gale.com/apps/doc/A618764393/HRCA?u=anon~90c40cbf&sid=googleScholar&xid=7031e225. Accessed 28 Jan. 2023.

Cho, D. (2007). The role of high school performance in explaining women's rising college enrollment. *Economics of Education Review, 26*, 450-462.

Choi, N., Chang, M., Kim, S., & Reio, T.G. (2015). A structural model of parent involvement with demographic and academic variables. *Psychology in the Schools, 52*, 154-167.

Cole, J., Kennedy, M., & Ben-Avie, M. (2009). The role of precollege data in assessing and understanding student engagement in college. *New Directions for Institutional Research, 2009*(141), 55-69.

Chute, E. (2006, April 2). Home and away: College-bound kids don’t stray far from home. *Pittsburgh Post-Gazette*. http://www.post-gazette.com/pg/06092/678836-298.stm

Cirelli, J., Konkol, A. M., Aqlan, F., & Nwokeji, J. C. (2018). Predictive analytics models for student admission and enrollment. *Proceedings of the International Conference on Industrial Engineering and Operations Management, 2018*(SEP), 1395-1403.

Cohen G. L., Garcia J., Purdie-Vaughns V., Apfel N., & Brzustoski P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science, 324*, 400-403.

Conley, A. M. (2012). Patterns of motivation beliefs: Combining achievement goal and expectancy-value perspectives. *Journal of Educational Psychology, 104,* 32-47.

Cornell University Division of Planning and Budget (2006). *Undergraduate enrollment trends, Fall 2006*. https://dpb.cornell.edu/documents/1000378.pdf.

Dara, S., Dhamercherla, S., Jadav, S. S., Babu, C. M., & Ahsan, M. J. (2022). Machine learning in drug discovery: A review. *Artificial intelligence review, 55*(3), 1947–1999.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves [Paper presentation]. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*, 498-506.

Dellas, H., & Sakellaris, P. (2003). On the cyclicality of schooling: theory and evidence, *Oxford Economic Papers, 55*(1), 148–172.

Deil-Amen, R., & Turley, R. L. (2007). A review of the transition to college literature in sociology. *Teachers College Record, 109*(10), 2324-2366.

DesJardins, S., and Gonzalez, J. (2002). Artificial neural networks: A new approach for predicting application behavior. *Research in Higher Education, 43*(2), 235-258.

Dorn, E., Dua, A., Law, J., & Ram, S. (2020). Higher education enrollment: Inevitable decline or online opportunity? McKinsey & Company. https://www.mckinsey.com/industries/public-and-social-sector/our-insights/higher-education-enrollment-inevitable-decline-or-online-opportunity.

Dornbusch, S., Ritter, P., Leiderman, P., Roberts, D., & Fraleigh, M. (1987). The relation of parenting style to adolescent school performance. *Child Development, 58*, 1244-1257.

Drake, B. M. and Walz, A. (2018), Evolving business intelligence and data analytics in higher education. *New Directions for Institutional Research, 2018*, 39-52.

Dreiseitl, S. & Ohno-Machado, L. (2002) Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics, 35*, 352-359.

Duncheon, J. C. (2015). The problem of college readiness. In W. G. Tierney & J. C. Duncheon (Eds.), *The problem of college readiness* (pp. 3-44). Albany: State University of New York Press.

Ekowo, M., & Palmer, I. (2016). *The promise and peril of predictive analytics in higher education: A landscape analysis.* New America. https://files.eric.ed.gov/fulltext/ED570869.pdf

Engberg, M. E., & Wolniak, G. C. (2010). Examining the effects of high school contexts on postsecondary enrollment. *Research in Higher Education*, *51*(2), 132-153.

Ewing, K.M., Beckert, K.A., & Ewing, B.T. (2010). The response of US college enrollment to unexpected changes in macroeconomic activity. *Education Economics*, *18*, 423 - 434.

Fair Test. (2022). Test-optional growth chronology 2005-2022. Retrieved from https://www.fairtest.org/sites/default/files/Optional-Growth-Chronology.pdf

Farrell, P. L., & Kienzl, G. S. (2009). Are state non-need, merit-based scholarship programs impacting college enrollment? *Education Finance and Policy*, *4*(2), 150-174.

Fawcett, T. (2006). An Introduction to ROC analysis. *Pattern Recognition Letters, 27,* 861-874.

Fedorovici, L. O., & Dragan, F. (2011). A comparison between a neural network and a SVM and Zernike moments based blob recognition modules [Paper presentation]. 6th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania (pp. 253-258).

Flashman, J. (2013). A cohort perspective on gender gaps in college attendance and completion. *Research in Higher Education, 54*(5), 545-70.

Fortin, N. M., Oreopoulos, P., & Phipps, S. (2015). Leaving boys behind gender disparities in high academic achievement. *Journal of Human Resources, 50*(3), 549–79.

Fraysier, K., Reschly, A., & Appleton, J. (2020). Predicting postsecondary enrollment with secondary student engagement data. *Journal of Psychoeducational Assessment*, *38*(7), 882–899.

Friedman, J. H, Kohavi, R. & Yun, Y. (1996). Lazy decision trees [Paper presentation]. In Proceedings of the 13th National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference (pp. 717-724), AAAI 96, IAAI 96 AAAI Press/ The MIT Press.

Fu, G., Yi, L., & Pan, J. (2018). Tuning model parameters in class-imbalanced learning with precision-recall curve. *Biometrical Journal, 61*, 652-664.

Fung, C. Y., & Adams, E. A. (2017). What Motivates Student Environmental Activists on College Campuses? An In-Depth Qualitative Study. *Social Sciences, 6*, 134.

Garrett, R. (2022, August 2). Inflation nation: The latest data on college enrollment in today’s on-edge economy. Encoura. https://encoura.org/inflation-nation-the-latest-data-on-college-enrollment-in-todays-on-edge-economy/

Gerasimovic, M., & Bugaric, U. (2018). Enrollment management model: Artificial neural networks versus logistic regression. *Applied Artificial Intelligence*, *32*(2), 153-164.

Geron, A. (2017). *Hands-On Machine Learning with Scikit-Learn & TensorFlow: Concepts, tools, and techniques to build intelligent systems*. Beijing: O`Reilly.

Gladiuex, L. E. (2004). America's untapped resource: Low-income students and the affordability of higher education. In R D. Kahlenberg (Ed.), *America's untapped resource* (pp. 17-58). New York: The Century Foundation Press.

Gomes, C. M. A., & Almeida, L. S. (2017). Advocating the broad use of the decision tree method in education. *Practical Assessment, Research, and Evaluation, 22*(10), 1-10.

Gonzalez, J., & DesJardins, S. L. (2002). Artificial neural networks: A new approach for predicting college application behavior. *Research in Higher Education, 43*(2), 235-258.

Goyette, K. A. (2008). College for some to college for all: Social background, occupational expectations, and educational expectations over time. *Social Science Research, 37*(2), 461–484.

Green, S. and Salkind, N. (2011). *Using SPSS for windows and Macintosh: Analyzing and understanding data*. Boston, MA: Prentice Hall.

Greene, B. A., & DeBacker, T. K. (2004). Gender and orientations toward the future: Links to motivation. *Educational Psychology Review, 16,* 91–120.

Gross, J. J. (2015). Emotion regulation: Current status and future prospects, *Psychological Inquiry, 26*(1), 1-26.

Guyon, I., Aliferis, C., & Elisseeff, A. (2007). Causal feature selection. In H. Liu & H. Motoda (Eds.), *Computational Methods of Feature Selection* (pp. 63–82), New York: Chapman and Hall/CRC.

Handwerk, M. L., Huefner, J. C., Ringle, J. L., Howard, B. K., Soper, S. H., Almquist, J. K., & Chmelka, M. B. (2008). The role of therapeutic alliance in therapy outcomes for youth in residential care. *Residential Treatment for Children & Youth, 25,*145–165.

Hansen. L, & Sargent, T. J. (2001). Robust Control and Model Uncertainty. *American Economic Review, 91*(2), 60-66.

Hanson, K., & Litten, L. (1982). Mapping the road to academia: A review of research on women, men, and the college selection process. In Perun, N.P., (Ed.), *The Undergraduate Woman, Issues in Education* (pp. 73-98), Lexington, KY: Lexington Books.

Harrell, F. E., Lee, K., & Mark, D. B. (1996). Tutorial in biostatistics multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine, 15,* 361-387.

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine, 15*(4), 361-387.

Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research, 5*, 1391–1415.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning; Data Mining, Inference and Prediction* (*2nd ed.),* Springer.

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-a-vis regression. *New Directions for Institutional Research, 2006*, 17-33.

Hetzel, R. L. (2009). Monetary Policy in the 2008-2009 Recession. *FRB Richmond Economic Quarterly, 95*(2), 201-233.

Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology, 6*, 25–45.

Holland, J. L., & Richards, J. M. (1965*). Academic and non-academic accomplishment-correlated or uncorrelated?* Iowa City, IA: American College Testing Program.

Hooker, S., & Brand, B. (2010). College knowledge: A critical component of college and career readiness. *New Directions for Youth Development*, *2010*(127), 75–85.

Horn, L., & Bobbitt, L. (2000). *Mapping the road to college: First-generation students' math track, planning strategies, and context of support.* Washington, DC: National Center for Education Statistics.

Horvat, E.M., Weininger, E.B., & Lareau, A. (2003). From social ties to social capital: Class differences in the relationships between schools and parent networks. *American Education Research Journal, 40*(2), 319-351.

Hosmer Jr., D.W., Lemeshow, S., & Sturdivant, R.X. (2013). *Applied Logistic Regression. (3rd ed.)*. Hoboken, NJ: Wiley.

Hossler, D. (1999). Effective admissions recruitment. *New Directions for Higher Education, 1999*(108), 15-30.

Hossler, D. (2002). The role of financial aid in enrollment management. *New Directions For Student Services, 2000*(89), 77-90.

Hossler, D., & Bean, J.P. (1990). *The strategic management of college enrollments.* San Francisco: Jossey-Bass.

Hossler, D., & Gallagher, K. S. (1987). Studying college choices: A three-phase model and the implication for policymakers, *College and University, 2,* 207-221.

Hossler, D., Lund, J. P., Ramin, J., Westfall, S., & Irish, S. (1997). State funding for higher education: The Sisyphean Tank. *The Journal of Higher Education, 68*(2), 160–190.

Hossler, D., & Palmer, M. (2008). Why understand research on college choice? In *Fundamentals of College Admission Counseling: A Textbook for Graduate Students and Practicing Counselors* (*2nd ed.*). Arlington, VA: Kendall Hunt Publishing.

Hoxby, C. M. (2009). The challenging selectivity of American colleges. *Journal of Economic Perspectives, 23*(4), 95-118.

Integrated Postsecondary Education Data System (2019). Price of attendance for full-time undergraduate students. https://nces.ed.gov/ipeds/datacenter/SelectVariables.aspx?stepId=2.

Jackson, G. (1978). Financial aid and student enrollment. *Journal of Higher Education, 49*, 548-78

Jackson, G.A. (1986). *Workable, comprehensive models of college choice. Final and technical report*, Carnegie Foundation for the Advancement of Teaching, National Institute of Education and Spencer Foundation, Washington, DC and Chicago, IL

Jadhav, S.D., & Channe, H.P. (2016). Efficient recommendation system using decision tree classifier and collaborative filtering. *International Research Journal of Engineering and Technology, 3*, 2113–2118.

James, G., Witten, D., Hastie, T., & Tibshirani. R. (2013). *An Introduction to Statistical Learning: With Applications in R*. New York, NY: Springer.

Japkowicz, N. (2013). Assessment metrics for imbalanced learning. In H. Haibo, & M. Yunqian (Eds.), *Imbalanced learning: Foundations, algorithms, and applications* (pp. 187-210). Wiley IEEE Press

Jarsky, K. M., McDonough, P. M., & Núñez, A. (2009). Establishing a college culture in secondary schools through P-20 collaboration: A case study. *Journal of Hispanic Higher Education, 8*(4), 357-373.

Jijo, B. T., & Abdulazeez, A. M. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends, 2*(1), 20-28.

Joseph, M., & Joseph, B. (2000). Indonesian Students’ Perceptions of Choice Criteria in the Selection of a Tertiary Institution: Strategic Implications. *International Journal of Educational Management, 14,* 40-44.

Kahlenberg, R D. (2004). *America's untapped resource: Low-income students in higher education (Ed.)*. New York: The Century Foundation.

Kao, G., & Thompson, J. S. (2003). Racial and ethnic stratification in educational achievement and attainment. *Annual Review of Sociology, 29*, 417-442.

Kao, G., & Tienda, M. (1998). Educational aspirations of minority youth. *American Journal of Education, 106*(3), 349–384.

Khouli, R. H., Macura, K. J., Barker, P. B., Habba, M. R., Jacobs, M. A., & Bluemke, D. A. (2009). The relationship of temporal resolution to diagnostic performance for dynamic contrast enhanced (DCE) MRI of the breast. *Journal of Magnetic Resonance Imaging, 30*(5), 999-1004.

Kim, D. (2004). The effect of financial aid on students' college choice: Differences by racial groups. *Research in Higher Education, 45,* 43–70.

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, 558-569.

Kinzie, J., Palmer, M., Hayek, J. C., Hossler, D., Jacob, S. A., and Cummings, H. (2004). *Fifty Years of College Choice: Social, Political and Institutional Influences on the Decision-Making Process*. Indianapolis, IN: Lumina Foundation for Education.

Klaauw, W. V. D. (2002). Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression–Discontinuity Approach. *International Economic Review, 43*(4), 1249-1287.

Klasik, D. (2012). The college application gauntlet: A systematic analysis of the steps to four-year college enrollment. *Research in Higher Education, 53,* 506-549.

Kleinfeld, J. (2009). No map to manhood: Male and female mindsets behind the college gender gap. *Gend. 26*, 171.

Kohl, M. (2012). Performance measures in binary classification. *International Journal of Statistics in Medical Research, 1*, 79-81.

Kotsiantis, S. B. (2013). Decision trees: a recent overview, *Artificial Intelligence Review, 39*, 261-283.

Knight, M., & Marciano, J. (2013). *College ready: Preparing Black and Latina/o youth for higher education – A culturally relevant approach.* New York, NY: Teachers College Press.

Knight, M. G., Marciano, J. E., Wilson, M., Jackson, I., Vernikoff, L., Zuckerman, K. G., & Watson, V. W. M. (2019). "It's all possible": Urban educators' perspectives on creating a culturally relevant, schoolwide, college-going culture for Black and Latino male students. *Urban Education, 54*(1), 35–64.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection [Paper presentation]. In Proceedings of the 14th International Joint Conference on Artificial Intelligence Volume 2, San Francisco, CA, USA (pp. 1137–1143). Morgan Kaufmann Publishers Inc.

Kohavi, R., & Provost, F. (1998). Glossary of terms. Machine Learning—Special Issue on Applications of Machine Learning and the Knowledge Discovery Process. *Machine Learning, 30*, 271-274.

Kohn, M. G. Manski, C. F., & Mundel, D. S. (1974). *An empirical investigation of Factors which influence college-going behavior.* Santa Monica, CA: The Rand Corporation.

Koshal, R. K., & Koshal, M. (2000). State appropriation and higher education tuition: What is the relationship? *Education Economics, 8*(1), 81-89.

Kotler, P. (1976). Applying marketing theory to college admissions. In *College entrance examination board, a role for marketing in college admission* (pp. 54-72). New York, NY: College Entrance Examination Board

Kotz, D. M. (2009). The financial and economic crisis of 2008: A systemic crisis of neoliberal capitalism. *Review of Radical Political Economics, 41*(3), 305–317.

Kowalik, E. (2011). Engaging alumni and prospective students through social media. In L. Wankel & C. Wankel (Eds.), *Higher education administration with social media* (pp. 211–227). United Kingdom: Emerald Group Publishing.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling (1st ed.)*. New York, NY: Springer.

Kunter, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models,* New York, NY: McGraw-Hill Irwin.

Lau, E.T., Sun, L. & Yang, Q. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Applied Sciences, 1*, 982.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path from Insights to Value. *MIT Sloan Management Review, 52*, 2.

Liu, W., Wu, U., Gao, X., & Feng, K. (2017). An early warning model of student achievement based on decision trees algorithm [Paper presentation]. 2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering, Hong Kong, China (pp. 517-222).

Luan, J., & Zhao, C. (2006). Practicing data mining for enrollment management and beyond. *New Directions for Institutional Research, 2006*, 117-122.

Ludemann, L., Grieger, W., Wurm, R., Wust, P., & Zimmer, C. (2006). Giloma assessment using quantitative blood volume maps generated by T1-weighted dynamic contrast-enhanced magnetic resonance imaging: a receiver operating characteristic study. *Acta Radiologica, 47,* 303-310.

Lux, T., Pittman, R., Shende, M., & Shende, A. (2016). Applications of supervised learning techniques on undergraduate admissions data. *In Proceedings of the ACM International Conference on Computing Frontiers (CF '16). Association for Computing Machinery, New York, NY, USA*, 412–417.

Massa, R. J., & Parker, A. S. (2007). Fixing the net tuition revenue dilemma: The Dickinson college story. *New Directions for Higher Educatio*n, 2007, 87-98.

McDonough, P. M. (1997). *Choosing colleges: How social class and schools structure opportunity*. State University of New York Press, Albany: NY.

Mehta, C. R., & Patel, N. R. (1995). Exact logistic regression: Theory and examples. *Statistics in Medicine. 14*(19), 2143-2160.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*(3)*,* 355–383.

Mikołajczyk, A. & Grochowski, M. (2018). Data augmentation for improving deep learning in image classification problem [Paper presentation]. In 2018 International Interdisciplinary PhD Workshop, Świnouście, Poland (pp. 117–122). IEEE.

Moraes, R., Valiati, J. F., & Neto, P. G. (2013), Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, *40*(2), 621-633.

Morgan, D. L. (1997). *Focus groups as qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.

Muschelli, J. (2020). ROC and AUC with a Binary Predictor: A potentially Misleading Metric. *Journal of Classification, 37*(3), 696-708.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data, 2,* 1-21.

National Bureau of Economic Research. (2008). *Business cycle dating committee announcement December 1, 2008*. https://www.nber.org/sites/default/files/2021-03/dec2008.pdf

National Center for Education Statistics. (2020). *Projections of education statistics to 2028.* https://nces.ed.gov/pubs2020/2020024.pdf

Nguyen, Q. H., Ly, H., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., Prakash, I., & Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Hindawi Mathematical Problems in Engineering, 2021*, 1-15.

Nuñez, A. M., & Kim, D. (2012). Building a multicontextual model of Latino college enrollment: Student, school, and state-level effects. *The Review of Higher Education, 35*(2), 237-263.

Okada, Y., Matsuyama, T., Morita, S., Ehara, N., Miyamae, N., Jo, T., Sumida, Y., Okada, N., Watanabe, M., Nozawa, M., Tsuruoka, A., Fujimoto, Y., Okumura, Y., Kitamura, T., Liduka R., & Ohtsuru, S. (2021). Machine learning-based prediction models for accidental hypothermia patients. j intensive care, *Journal of Intensive Care, 2021*(9), 6.

Obuchowski, N.A. (2003). Receiver operating characteristics curves and their use in radiology. *Radiology, 229*, 3-8.

Olson, D. L., and Delen, D. (2008). *Advanced data mining techniques*. Heidelberg, German: Springer.

Parikh, K. S., & Shah, T. P. (2016). Support Vector Machine – a Large Margin Classifier to Diagnose Skin Illnesses. *Procedia Technology, 23,* 369-375.

Parker, B.J., Günter, S. & Bedo, J. (2007). Stratification bias in low signal microarray studies. *BMC Bioinformatics, 8*, 326.

Paulsen, M. B., & St. John, E. P. (2002). Social class and college costs: Examining the financial nexus between college choice and persistence. *The Journal of Higher Education, 73*, 189-236.

Perna, L. W. (2005). The benefits of higher education: Sex, racial/ethnic, and socioeconomic group differences. *The Review of Higher Education, 29*(1), 23-52.

Perna, L. W. (2006). Understanding the relationship between information about college prices and financial aid and students' college-related behaviors. *American Behavioral Scientist, 49*(12), 1620–1635.

Perna, L. W., & Titus, M. A. (2005). The Relationship between Parental Involvement as Social Capital and College Enrollment: An Examination of Racial/Ethnic Group Differences, *The Journal of Higher Education, 76*(5), 485-518, Peruta, A., & Shields, A. B. (2018). Marketing your university on social media: a content analysis of Facebook post types and formats. *Journal of Marketing for Higher Education, 28*(2), 175-191.

Pickett, W. L. (1972). *Techniques of Institutional Research and Long Range Planning for Colleges and Universities, Volume I: Enrollment Projections, Induced Course Load Matrix, Faculty Planning*. Kansas City, MO.: Midwest Research Institute, Economics and Management Science Division.

Platt, J. (2000). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers*. Cambridge: MIT Press.

Pochet, N., & Suykens, J. (2006). Support vector machines versus logistic regression: Improving prospective performance in clinical decision-making. *Ultrasound Obstet Gynecol, 2006,* 607-608.

Porter, A. C., & Polikoff, M. S. (2012). Measuring Academic Readiness for College. *Educational Policy, 26*(3), 394–417.

Psacharopoulos, G. (1973) *Returns to Education: An International Comparison*. San Francisco, CA: Elsevier, Jossey-Bass.

Radford, N. (1996). *Bayesian learning for neural networks.* New York, NY: Springer

Ragab, A. M., Noaman, A. Y., Al-Ghamdi, A. S. and Madbouly, A. I. (2014, June 19). *A comparative analysis of classification algorithms for students college enrollment approval using data mining* [Conference presentation]. Interaction Design in Educational Environments (IDEE), Albacete, Spain. https://www.semanticscholar.org/paper/A-Comparative-Analysis-of-Classification-Algorithms-Ragab-Noaman/4798801dca97fb6e1cf127c12f049f92d3757a2c

Rahman, N. & Iverson, S. (2015). Big data business intelligence in bank risk analysis. *International Journal of Business Intelligence Research, 6*(2), 55-77.

Raileanu, L.E., & Stoffel, K. (2004). Theoretical Comparison between the Gini Index and Information Gain Criteria. *Annals of Mathematics and Artificial Intelligence, 41*, 77-93.

Ranganathan, P., Pramesh, C. S., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. *Perspectives in clinical research, 8*(3), 148–151.

Ren, J. (2012). ANN vs. SVM: Which one performs better in classification of MCCs in mammogram imaging. *Knowledge-Based Systems, 26,* 144-153.

Reynolds J. R., & Burge S. W. (2008). Educational expectations and the rise in women's educational attainments. *Social Science Research, 37*(2), 485–499.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. & Chica-Rivas, M. (2015) Machine Learning Predictive Models for Mineral Prospectivity: An Evaluation of Neural Networks, Random Forest, Regression Trees and Support Vector Machines. *Ore Geology Reviews, 71*, 804-818.

Rodríguez-Muñiz L. J., Bernardo, A.B., Esteban, M., & Díaz, I. (2019). Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques? *PLOS ONE, 14*(6), e0218796. https://doi.org/10.1371/journal.pone.0218796

Roscigno, V. J. (2000). Family/school inequality and African-American/Hispanic achievement. *Social Problems, 47*(2), 266–290.

Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences, 3*(2), Article 272.

Sammut, C. & Webb, G. I. (2017). *Encyclopedia of machine learning and data mining*, Germany: Springer.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*, 160.

Sawiris, M. Y. (1970). The project of college enrollment. *Multivariate Behavior Research, 5*(1), 83-100.

Seger, C. (2018). An investigation of vategorical variable encoding techniques in machine learning: Binary versus one-hot and feature hashing.

Senaviratna, N., & Cooray, T. (2019). Diagnosing multicollinearity of logistic regression model. *Asian Journal of Probability and Statistics, 5*(2), 1-9.

Seres, L., Pavlicevic, V., Tumbas P. (2018, March 5-7) Digital transformation of higher education: competing on analytics [Paper presentation]. Proceedings of International Technology, Education and Development (INTED) Conference (pp. 9491-9497). Subotica, Serbia.

Shabestari, S. S. & Herzog, M. & Bender, B. (2019). A survey on the applications of machine learning in the early phases of product development. *Proceedings of the Design Society: International Conference on Engineering Design, 1*, 2437-2446.

Shah, S., & Sastry, P. S. (1999). New algorithms for learning and pruning oblique decision trees. *IEEE Transactions on Systems, Man, and Cybernetics*, *Part C: Applications and Reviews 29*(4), 494–505.

Sharma, D. & Kumar, N. (2017). A review on machine learning algorithms, tasks and applications. *International Journal of Advanced Research in Computer Engineering & Technology, 6*, 10.

Slim, A., Heileman, G. L., Kozlick, J., & Abdallah, C. T. (2014, December). Predicting student success based on prior performance [Paper presentation]. In Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on (pp. 410-415). IEEE.

Solis, A. (2017). Credit access and college enrollment. *Journal of Political Economy, 125*(2), 124-147.

Sollich, P. (2002). Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine Learning, 46*, 21–52.

Song, Q., & Chisson. B. S. (1993). Fuzzy time series and its models. *Fuzzy Sets and Systems, 54*(3), 269-277.

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry, 27*(2), 130–135.

St. John, E.P., Hu, S., & Fisher, A.S. (2010). *Breaking through the access barrier: How academic capital can improve policy in higher education.* New York: Routledge.

St. John, E. P., & Noell, J. (1989). The effects of student financial aid on access to higher education: An analysis of progress with special consideration of minority enrollments. *Research in Higher Education, 30*(1), 563 -58.

Stage, F. K. & Hossler, D. (1989). Differences in family influences on college attendance plans for male and female ninth graders. *Research in Higher Education, 30*(3), 301-315.

Stange, K. (2012). An empirical investigation of the option value of college enrollment. *American Economic Journal: Applied Economics, 4*(1), 49-84.

Statnikov, A., Wang, L. & Aliferis, C.F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics 9*, 319.

Sternberg, R. J. (2010). WICS: A new model for school psychology. *School Psychology International, 31*(6), 599–616.

Stone, M. (1977). An asymptotic equivalence of choice model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 44-47.

Y. Tang, Y. -Q. Zhang, N. V. Chawla, & S. Krasser (2009), SVMs modeling for highly imbalanced classification, *IEEE Transactions on Systems, Man, and Cybernetics Part B, 39*(1), 281-288.

Tay, F. E. H., & Cao, L. J. (2002). Modified support vector machines in financial time series forecasting. *Neurocomputing, 48*, 847-861.

Teachman, J.D., & Polonko, K. D. (1988). Marriage, parenthood, and the college enrollment of men and women. *Social Forces, 67*(2), 512–523.

Thomas, D. M., Kuiper, P., Zaveri, H., Surve, A., & Cottam, D.R. (2017). Neural networks to predict long-term bariatric surgery outcomes. *Bariatric Times, 14*(12), 14–7.

Truman, E. M. (2003). The limits of exchange market intervention. In *Dollar overvaluation and the wold economy* (Ed.). C. Fred Bergsten and John Williamson. *Special Report 16*. Washington: Institute for International Economics.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology, 49*(11), 1225-1231.

Turley, R. N. L. (2009). College proximity: Mapping access to opportunity. *Sociology of Education, 82*(2), 126–146.

Turley, R. N. L., Santos, M., & Ceja, C. (2007). Social origin and college opportunity expectations across cohorts. *Social Science Research, 36*, 1200-1218.

Ustunner, M., Sanli, F. B., & Abdikan, S. (2016, July 12-19). Balanced vs imbalanced training data: Classifying rapideye data with support vector machines [Paper presentation]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLI-B7, 2016 XXIII ISPRS Congress, Prague, Czech Republic (pp. 379-384).

Vialardi, C., Bravo, J., Shafti, L., & Ortigosa, A. (2009, July 1-3*). Recommendation in higher education using data mining techniques* [Conference presentation]. International Conference on Educational Data Mining (EDM), 2nd, Cordoba, Spain. https://eric.ed.gov/?id=ED539088

Vialardi, C., Chue, J., Peche, J. P., Alvarado, G., Vinatea, B., Estrella, J., & Ortigosa, A. (2011). A data mining approach to guide students through the enrollment process based on academic performance. *User Modeling and User-Adapted Interaction, 21*, 217–248.

Vossensteyn, J. J. (2005). *Perceptions of student price-responsiveness: A behavioural economics exploration of the relationships between socio-economic status, perceptions of financial incentives and student choice.* [PhD Thesis - Research UT, graduation UT, University of Twente]. Center for Higher Education Policy Studies (CHEPS).

Walczak, S., & Sincich, T. (1999). A comparative analysis of regression and neural works for university admissions. *Information Sciences, 119* (1-2), 1-20.

Wang, J., Neskovic, P., & Cooper, L. N. (2005). Training Data Selection for Support Vector Machines. *International Conference on Natural Computation, 2005,* 554-564.

Warner, R. M. (2013). *Applied statistics: From bivariate through multivariate techniques (2nd ed.).* Sage Publications, Inc.

Wearne, E. (2018). Oakeshott, Schumacher, and an examination of the tension between the "college and career readiness" consensus and school choice in american education policy. *Policy Futures in Education, 16*(3), 291–305.

Western Interstate Commission of Higher Education. (2020). *Knocking at the college door projections of high school graduates.* Western interstate commission for higher education. https://files.eric.ed.gov/fulltext/ED610996.pdf.

Windolf, P. (1997). *Expansion and structural change: Higher education in Germany, the United States, and Japan.* Boulder, CO: Westview.

Wissmann, M., Shalabh, S., & Toutenburg, H. (2014). Role of categorical variables in multicollinearity in linear regression model. *Journal of Applied Statistical Science, 19*(1), 99-113.

Wong, T., & Yang. N, (2017). Dependency analysis of accuracy estimates in k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering, 29*(11), 2417-2427.

Yadav, S., & Shukla, S. (2016). Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification [Paper presentation], 2016 IEEE 6th International Conference on Advanced Computing (IACC), Bhimavaram, India (pp. 78-83).

Yang, F. J. (2019). An Extended Idea about Decision Trees. 2019 International Conference on Computational Science and Computational Intelligence, Las Vegas, NV, USA (pp. 349-354).

Yu, C. S., Lin, C. J., & Hwang, J. K. (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein science : a publication of the Protein Society*, *13*(5), 1402–1406.

Zeng, D., Gao, F., Hu, K., Jia, C., & Ibrahim, J. G. (2015). Hypothesis testing for two-stage designs with over or under enrollment. *Statistics in medicine*, *34*(16), 2417–2426.

Zeng, X., & Martinez, T. R. (2000) Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence, 12*(1), 1-12, Zullig, K. J., Koopman, T. M., Patton, J. M., & Ubbes, V. A. (2010). School climate: Historical review, instrument development, and school assessment. *Journal of Psychoeducational Assessment*, *28*(2), 139–152.

VITA

Dr. Kye graduated from Iowa State University in Ames, Iowa in 2011 with a Bachelor of Science in Statistics and a minor in Mathematics. In 2014 she earned a Master's degree from the University of Massachusetts – Amherst, majoring in Statistics. In 2016 she also earned another Master's degree from the University of Iowa, majoring in Mathematics Education. Dr. Kye began her doctoral studies in the Research Methodology program at Loyola University Chicago in Fall of 2018. Dr. Kye is a senior research & strategy analyst with more than six years of experience in quantitative modeling and consulting in the higher education sector.