Summer 2020

Development of a system architecture for the prediction of student success using machine learning techniques

Tatiana A. Cardona

Department: Engineering Management and Systems Engineering

DEVELOPMENT OF A SYSTEM ARCHITECTURE FOR THE PREDICTION OF STUDENT SUCCESS USING MACHINE LEARNING TECHNIQUES

by

TATIANA ALEJANDRA CARDONA SEPULVEDA

A DISSERTATION

Presented to the Graduate Faculty of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

in

SYSTEMS ENGINEERING

2020

Approved by:

Elizabeth Cudney, Advisor
Cihan H. Dagli
Benjamin Kwasa
Susan L. Murray
Douglas K. Ludlow

PUBLICATION DISSERTATION OPTION

This dissertation consists of the following six articles, formatted in the style used by the Missouri University of Science and Technology:

Paper I, found on pages 6–39, has been submitted to Journal of College Student Retention: Research, Theory & Practice in Oct 2019 and is under revision.

Paper II, found on pages 40-66, is intended for submission to *Expert Systems Journal.*

Paper III, found on pages 67-80, has been published in the proceedings of the ASEE Annual Conference & Exposition, Tampa, FL, in Jun 2019.

Paper IV, found on pages 81-96, has been published in the proceedings of IISE Annual Conference, Orlando, FL, in May 2019.

Paper V, found on pages 97-110, has been published in the proceedings of the International Conference on Production Research Manufacturing Innovation: Cyber Physical Manufacturing, Chicago, IL, in August 2019.

Paper VI, found on pages 111- 129, has been submitted to the proceedings of ASEE Annual Conference & Exposition, virtual conference, June 2020 and was accepted for publication.

ABSTRACT

The goals of higher education have evolved through time based on the impact that technology development and industry have on productivity. Nowadays, jobs demand increased technical skills, and the supply of prepared personnel to assume those jobs is insufficient. The system of higher education needs to evaluate their practices to realize the potential of cultivating an educated and technically skilled workforce. Currently, completion rates at universities are too low to accomplish the aim of closing the workforce gap. Recent reports indicate that 40 percent of freshman at four-year public colleges will not graduate, and rates of completion are even lower for community colleges. Some efforts have been made to adjust admission requirements and develop systems of support for different segments of students; however, completion rates are still considered low. Therefore, new strategies need to consider student success as part of the institutional culture based on the information technology support. Also, it is key that the models that evaluate student success can be scalable to other higher education institutions. In recent years machine learning techniques have proven to be effective for such purpose. Then, the primary objective of this research is to develop an integrated system that allows for the application of machine learning for student success prediction. The proposed system was evaluated to determine the accuracy of student success predictions using several machine learning techniques such as decision trees, neural networks, support vector machines, and random forest. The research outcomes offer an important understanding about how to develop a more efficient and responsive system to support students to complete their educational goals.

ACKNOWLEDGMENTS

I would like to express my special appreciation and thanks to my advisor, Dr. Elizabeth Cudney, for providing me guidance, for her enthusiasm for the project, and for her support, encouragement and patience.

I would also like to thank my committee members, Dr. Cihan Dagli, Dr. Benjamin Kwasa, Dr. Susan Murray, and Dr. Douglas Ludlow, for letting my defense be a gratifying moment, and for their brilliant comments and suggestions, thanks to them.

I also wish to thank Dr. Suzanna Long who provided great support during the time of my PhD studies. Thanks also to the department staff who have been very diligent and caring.

A special thanks to my family for all the support they have shown. Words cannot express how grateful I am to my mother for all the sacrifices that she has made on my behalf, and my uncle for making it possible to start this important journey.

Thanks to my friends you for supporting me and encouraging me throughout this experience.

I thank God for getting me through all these experiences with perseverance and love.

TABLE OF CONTENTS

SECTION

LIST OF ILLUSTRATIONS

PAPER I Page

PAPER II

PAPER III

PAPER IV

PAPER V

PAPER VI

LIST OF TABLES

PAPER I Page

PAPER II

PAPER III

PAPER VI

PAPER V

PAPER VI

1. INTRODUCTION

1.1. BACKGROUND AND MOTIVATION

The goals of access to higher education have evolved through time based on the impact that technology development and industry have on U.S. productivity (Handel, 2013). Around the mid-20's the main goal was to increase access capacity (Bailey, 2017). But at the end of the century, research indicated there was a skill gap in our workforce, data showed that job demand increased requirements in technical skills and more advanced degrees (Carnevale et al., 2016), and there was a limited number of qualified applicants (Restuccia & Taska, 2018). Therefore, the government, universities, and community colleges started to turn their attention towards student success and, consequently, student completion rates and retention (Matthews, 2012).

The National Student Clearinghouse Research Center (Shapiro et al., 2018) indicates that 65.7% of students at four-year public institutions graduate within six years, while the number decreases dramatically to 39.2% of students in two-year public institutions graduate within three years. According to Kirp (2019), 40% of college freshman will not graduate. The rise in students attending community colleges, or two-year institutions, during their first two years has resulted in a need to include this important aspect of the ecosystem in the analysis. Further investigation of graduation data indicates that nationally only 17% of students that start at a community college will transfer and graduate from a four-year institution within six years (Jenkins & Fink, 2015). The reasoning behind these statistics is multifaceted and will not be solved without concerted effort from all partners in higher education.

The system of higher education needs to evaluate their practices to realize the potential of cultivating an educated and technically skilled workforce. Given the disparate outcomes at institutions, increasing student retention in higher education is of important interest as it reflects institutional commitment to the students (Williford & Schaller, 2005). The current literature suggests that an important reason for failing to improve student success is the scope of the investigation and implementation of possible solutions (Kelly & Schneider, 2012). Institutions have concentrated their efforts on studying small segments of students such as minority, low income students, first generation college students, and freshmen, among other groups (Governor's Bussiness Council, 2002; Alkhasawneh & Hargraves, 2014; Márquez-Vera et al., 2016; Thomas & Teras, 2014; Iam-On & Boongoen, 2017; Kondo et al., 2017). Although imperative insights can be obtained from such studies, and they represent benefits for each specific segment, the results for the broader view have not been very promising (Bahar & Eylem, 2015). The implementation of reforms and strategies should be done in a progressive and broader manner to provide evidence of improvement of student completion (Bailey et al., 2005). To support this perception, three major factors come into play (Bailey et al., 2015; Hiles, 2017; Grajek, 2017; Bailey, 2017; Klempin & Karp, 2018). The first factor is the culture of student success. Reforms need to take place to integrate the system forces (e.g., management, administrative stuff, and faculty) to make student success a priority and intrinsic part of an educational institution's strategies and activities. Second, information technology (IT) should be recognized as an important agent for improving student success. And third, the scalability, results must be scalable, and the applications must be able to be successfully applied in other institutions.

Therefore, to develop reforms and strategies to improve student success, mechanics of the system must be identified in order to recognize the different ways student retention can be evaluated. Also, the institution should determine the interactions that systematically impact student success.

In recent years machine learning techniques have been applied to analyze student data, which aligns with the focus on improving the processing of information through data mining (Cardona et al., 2019a), and can help achieve scalability (Dahlstrom, 2016). According to the literature, techniques such as decision trees, neural networks, and support vector machines, offer predictions of student dropout with high confidence (Pereira & Zambrano, 2017). These techniques are tools that also help to determine the factors that influence student retention and completion rates.

The creation of a system that allows for predictive models that help in the recognition of students at risk for attrition, will enable timely interventions. Universities and community colleges can provide intentional student advising and planning. Further, higher education institutions can develop retention strategies that focus on identifying student needs that meet their specific campus needs (Slim et al., 2014).

1.2. AIMS AND APPROACHES

The primary objective of this dissertation is to offer a systematic model to establish the agents that intervene in student success, not as separated aspects but as an integrated system. To accomplish this, the proposed system and its interactions will be tested using machine learning techniques to determine if they can produce accurate student success predictions. The prediction models will be applied to university and community college data.

Therefore, the major contributions of this research can be summarized as follows:

- Investigate in the literature the performance of prediction models for student success of the different machine learning techniques that have been applied and identify the variables that have had a high impact on the models.
- Formulate a complex system structure that allows for the evaluation of strategies to improve student success in higher education specifically community college and university environments. The model will represent the information flow of a student that enters the higher education system.
- Implement machine learning models such as neural networks, decision trees, support vector machines, and random forest techniques using the proposed system. The development of these models validates the system structure and allow for the identification of the impact of the variables on student success.

This modeling approach has an important role in the effective generation of variable-focused strategies for intentional advising.

1.3. DISSERTATION SYNOPSYS

This dissertation is organized as follows:

Section 1, provides an introduction. It briefly introduces the motivation of this research.

Section 2, presents a systematic literature review. It reviews the literature on the prediction of student retention in higher education through machine learning algorithms based on retention measures such as dropout risk, attrition risk, and completion risk.

Section 3 proposes the structure of a higher education system through the integration of factors that allow for the prediction of student success.

Section 4, presents the analysis of student data with the aim of predicting degree completion within three years for STEM community college students using decision trees, specifically Classification and Regression tree (C&RT).

Section 5, studies the application of neural networks (NN) to predict degree completion within three years by STEM community college students. This study enables the classification of the input variables into expected results, retention, and completion.

Section 6, presents the analysis of student data to predict degree completion within three years for STEM community college students using support vector machines (SVM), a machine learning technique.

Section 7, presents the analysis of student data with the aim of predicting degree completion within three years for STEM community college students using an ensemble machine learning technique, specifically random forest (RF).

PAPER

I. DATA MINING AND MACHINE LEARNING RETENTION MODELS IN HIGHER EDUCATION, A SYSTEMATIC REVIEW

Tatiana A. Cardona[a], Elizabeth A. Cudney[a]

[a]Deparment of Engineering Management and Systems Engineering, Missouri University of Science and Technology

ABSTRACT

This study presents a systematic review of the literature on the prediction of student retention in higher education thorough machine learning algorithms based on retention measures such as dropout risk, attrition risk, and completion risk. A systematic review methodology was employed that comprised of review protocol, requirements for study selection, and analysis of paper classification. This review aims to answer the following research questions: (1) what techniques are currently used to predict student retention rates and which have shown better performance under specific contexts?, (2) which factors influence the prediction of completion rates in higher education?, and (3) what are the challenges with the disposition of the results? Increasing student retention in higher education is critical as it increases graduation rates. Further, predicting student retention provides insight into opportunities for intentional student advising. This review provides a perspective on research related to the predicting student retention through machine learning.

1. INTRODUCTION

The United States has emerged from the Great Recession and there is a growing job surplus due to the limited number of qualified applicants for these jobs (Restuccia & Taska, 2018). Research indicates there is a skill gap in our workforce which will only continue to widen without corrective action in higher education. The jobs of today demand increased technical skills and more advanced degrees than in prior generations. When looking at the recovery data, the jobs that have filled the void are jobs that required a college degree while those without advanced training have continued to struggle (Carnevale et al., 2016). These factors propel the higher education ecosystem to turn inward and find solutions to some of the ailments that plague it such as increasing cost, inequity, retention, and completion rates. The National Student Clearinghouse Research Center (Shapiro et al., 2018) indicates that 65.7% of students at our-year public institutions graduate within six years, while the number decreases dramatically to 39.2% of students in two-year public institutions graduate within three years. According to Kirp (2019), 40% of college freshman will not graduate and, "Dropouts are nearly twice as likely as college grads to be unemployed, and they are four times more likely to default on student loans, thus wrecking their credit and shrinking their career options." The system of higher education will need to evaluate their practices to realize the potential of an education and technically skilled workforce. The rise in students attending community colleges, or two-year institutions, during their first two years has resulted in a need to include this important aspect of the ecosystem in the analysis. Further investigation of graduation data indicates that nationally only 16% of students that started at a community college will transfer and graduate from a four-year institution within six years. The reasoning behind these statistics is multifaceted and will not be solved without concerted effort from all partners in higher education.

According to Morris (2016), better data is needed in the decision-making process to improve student success. Almost 50 years ago the need for data analysis was recognized to answer important questions about student enrollment, faculty ranks and distribution and revenue and expenditures. However, nowadays the ability to make the data useful is not running at the same pace as the data collection and a significant amount of data is left without use. The U.S. Department of Education set a goal of preparing a society with individuals capable to "understand, explore and engage with the world", which are specific skills that can be achieved through STEM majors. Given the disparate outcomes at institutions, increasing student retention in higher education is of important interest as it reflects institutional commitment to the students (Williford & Schaller, 2005). Retention rates are one of the main concerns for universities and colleges, perhaps more important to community colleges due to this being a growing entry point for higher education (National Science Board, 2016; Chen, 2013; Hoffman et al., 2010), particularly with respect to STEM students (Snyder & Cudney, 2018). Students completing their degrees in the expected time directly impacts funding and the reputation of the institution, as it reflects institutional commitment with the educational goals.

In addition, determining the factors that influence student retention and completion rates provides insight into opportunities for intentional student advising, better planning, and development of retention strategies based on student needs (Slim et al., 2014). In recent years machine learning techniques have been applied to analyze student data, which aligns with the focus on improving the processing of information through data mining (Cardona et al., 2019a) using methods such as artificial neural networks (Cardona et al., 2019b) and support vector machines (Cardona & Cudney, 2019). According to the literature, those techniques offer predictions of student dropout with high confidence (Pereira & Zambrano, 2017).

This study presents a systematic review of the implementation of machine learning techniques to improve retention rates in educational institutions. This study aimed to answer the following questions:

What techniques are currently being used to predict student retention rates and which have shown better performance under specific contexts?

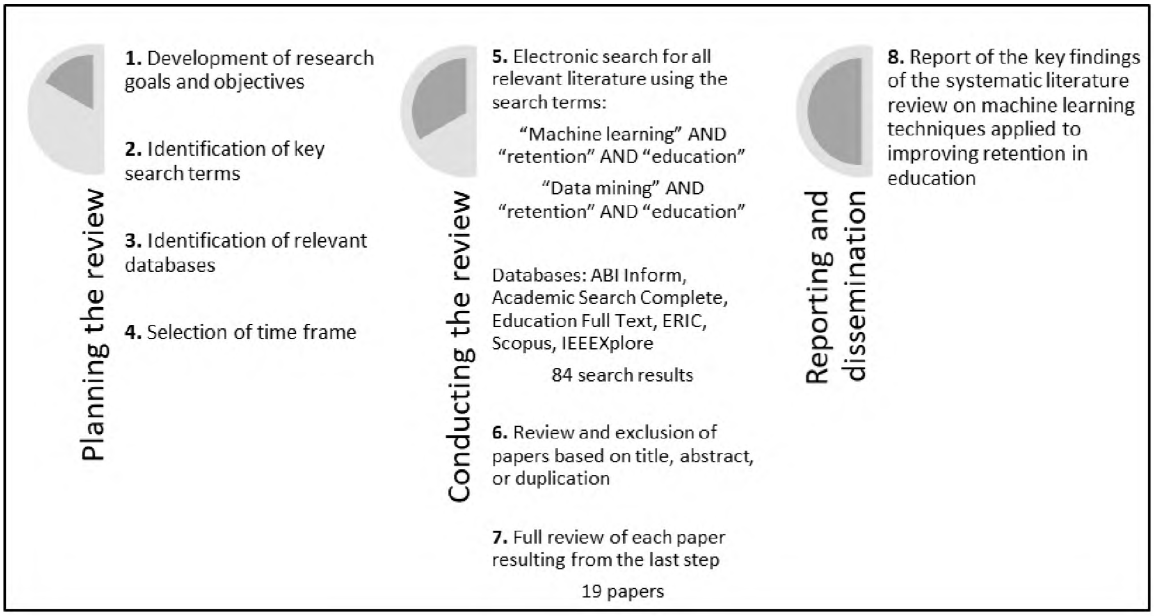
Which factors influence the prediction of completion rates in higher education and, what are the challenges with the disposition of the results.

A systematic literature review approach that was proposed by Tranfield et al. (2003) was employed to collect papers within the scope of this study. The studies were classified to determine the papers that would be further analyzed. The main characteristics evaluated were the techniques used for prediction and their performance along with the factors used in the models and the source of the information.

The structure of this paper is as follows. The next section contains the introduction of the research approach. In the third section, the application of machine learning techniques for the prediction of student retention is reviewed. Then the findings are analyzed and presented. Finally, concluding remarks and suggestions for future research are presented.

2. SYSTEMATIC REVIEW METHODOLOGY

The systematic review was developed in three stages as proposed by Tranfield et al. (2003). First, the planning process, followed by conducting the review, and finally reporting and dissemination. Each stage had several steps as illustrated in Figure 1.



Planning the review

1. Development of research goals and objectives

2. Identification of key search terms

3. Identification of relevant databases

4. Selection of time frame

Conducting the review

5. Electronic search for all relevant literature using the search terms:

"Machine learning" AND "retention" AND "education"

"Data mining" AND "retention" AND "education"

Databases: ABI Inform, Academic Search Complete, Education Full Text, ERIC, Scopus, IEEEXplore

84 search results

6. Review and exclusion of papers based on title, abstract, or duplication

7. Full review of each paper resulting from the last step

19 papers

Reporting and dissemination

8. Report of the key findings of the systematic literature review on machine learning techniques applied to improving retention in education

Figure 1. Stages of the systematic review

2.1. PLANNING THE REVIEW

The main objective of the systematic review was to identify and organize the available literature on the application of machine learning techniques to predict student retention rates. Further, the intent was to determine the relevant factors that have been used and recognized as important to predict student completion rates in higher education.

The key words "machine learning", "data mining", "retention" and "education" were used in the search. Articles published until August 31, 2018 that utilized machine learning techniques were used for this systematic review. Databases used in the search to ensure inclusion of the relevant literature were: *ABI Inform, Academic Search Complete, Education Full Text, ERIC, Scopus and IEEEXplore*. The selection of databases was based on the research domains and their types of publications to ensure representativeness of the available literature in terms of the systematic review objectives.

The search criteria for the literature selection include journals and peer reviewed publications, as well as articles published in English and Spanish. Books and non-referred publications were excluded. The relevant literature was organized according to the implementation of machine learning models for predicting student retention in higher education.

2.2. CONDUCTING THE REVIEW

The literature search was performed using the key words accompanied by the term 'AND'. Therefore, the search strings were "machine learning" AND "retention" AND "education" and "data mining" AND "retention" AND "education". In the field section, 'All text' was selected and literature was searched through the current date of the search, which was August 31, 2018.

A total of 87 results were obtained from the search process. Each paper was evaluated by title and abstract using the criteria specified in the planning stage section of this document. After applying the exclusion criteria and removing duplicates, only 19 papers remained for the full review in the last stage of the systematic search: reporting and dissemination. The remaining papers were reviewed to categorize the techniques used for the prediction of retention rates and the identification of the factors/variables used in the prediction models.

3. LITERATURE REVIEW OF PREDICTION MODELS FOR STUDENT RETENTION USING MACHINE LEARNING TECHNIQUES

Student retention and degree completion are directly related with university rankings. In fact, they are considered measures of institutional performance and success. Increasing retention and completion rates in higher education in the United States, specifically for STEM majors, is one of the objectives of the U.S. Department of Education. In these terms, the analysis of student data is vital to determine the factors that influence degree completion rates, providing an opportunity to investigate and improve intentional student advising. Recently, machine learning techniques have been applied to process educational data focused on student success measured as risk of dropout, attrition risk, and completion risk, which translates to retention and graduation rates (Williford & Schaller, 2005). This section provides a discussion of the studies that apply machine learning models for the prediction of retention or completion rates in higher education.

3.1. IMPLEMENTATION OF MACHINE LEARNING TECHNIQUES TO PREDICT DEGREE COMPLETION

McAleer and Szakas (2010) developed a model to predict retention risk from past data and determine if transfer students have a higher retention risk. Data from 10 years (1997-2007) was collected and used in this study. The prediction classes included student retained (persisting degree) and not retained, and the database included 13 variables. The methodology used Naïve Bayes and support vector machines (SVM). SVM obtained a 79.59% classification accuracy, which surpassed the results of the Naïve Bayes model (57.35%). The study also discovered that grades in 100 and 200 level courses are the most important variables for predicting retention. Further, age and gender were not determined to be relevant factors for retention. The research concluded that transfer students do not have increased retention risk.

Research by Delen (2010) used the cross industry standard process for data mining to predict and explain reasons for student attrition. The study is focused on retention prior to sophomore and the models presented had approximately 80% of accuracy. This study showed the individual application of several classification methods such as neural networks (NN), decision trees (DT) specifically the C5 algorithm, SVM, and logistic regression (LR). The results were compared to the use of different ensembles, which included 1.random forest (RF) which is an ensemble of several decision trees with sizes and variables chosen randomly for the sample, 2. boosted trees different from random forest in the way the new trees in the ensemble are generated from the residuals from the preceding tree , and 3. Information fusion which is the combination of different predictors. The dataset for analysis was composed of 16,066 students enrolled as freshmen during 2004 and 2008. The models were applied to the original dataset and later to a well-balanced dataset taken from original data but with equally represented classes to predict dropout. For individually applied techniques, the most accurate results were obtained when using the well-balanced dataset in all cases. The best results were from the SVM technique; however, using DT offered the advantage of a more transparent structure without significantly impacting accuracy. When using the ensembles with the well-balance data set, a slight improvement in the accuracy of the predictions was achieved. A sensitivity analysis showed the variables that impact at-risk student prediction for this study were student scholarships, loans, and fall GPA.

In a similar study, Delen (2011) compared three different prediction models for freshmen student attrition. The techniques used to develop the analytical models were NN, DT specifically C5 algorithm, and LR. Institutional data collected from eight years was used to develop the models. The research found that, with appropriate data and variables, machine learning techniques could predict student attrition with approximately 80% accuracy. NN obtained the best performance, although DT offered a more visual structure of the results. The classification of factors indicated that fall GPA, loans, and financial aid had a significant impact on predicting student attrition. In other words, educational and financial variables are important when predicting freshman attrition.

A student success system was developed by Essa and Ayad (2012), which provides an analytical platform for pre-emptively measuring student success. The system offers advanced data visualization for diagnostic measures and a case management tool for managing interventions. The visualization interface shows information in percentages for college preparedness, success index, attendance, completion, participation, social learning, actual grade, and prediction of grade. The model was created using healthcare models that predict patient risk level of disease. The student success system design uses data from operational sources such as the learning management system (LMS) and web logs, which are aggregated and stored using the extraction, transformation, lead (ETL) process. The data was captured every day for each student and processed using machine learning techniques to generate a prediction of dropout risk. The student success system interacts with the user through a mobile app or desktop browser. The system was developed with the aim of offering generalization of the results into different learning contexts such as different institutions, different courses, among others. However, the findings showed the applicability of the system to other institutions needed a great deal of customization, then it was presented as research limitation. The value of the system resides on the visualization of the data and information management interface provided, which was developed as to show the student status in four sub-categories: attendance, completion, participation and social learning. The final classification of the student into being at risk of dropout or not was made using an ensemble of different algorithms

In a related study, Slim et al. (2014) proposed a prediction model for students' success in their early academic career. Student success was measured using the GPA (letter and number) of previous courses. Bayesian belief network (BBN) technique was applied using a database of 115,746 students from the University of New Mexico. To test the predictions, information from an additional 400 students. Then, a simulation was created to empirically validate the implementation of the BBN. To develop the simulation, conditional probabilities were deployed, meaning the probability of having certain grade in class B depend of the student grade on class A that was pre-requisite of B, in this way the model will account for the dependencies and transitions from a certain grade to another.. The accuracy of the models was measured using the mean squared error (MSE) and margin error (percentage points of variation with actual population measure). It was possible to determine that the BBN had a good performance with a margin error of 0.16 (4.3 was the maximum GPA value that can be achieved). Future research was discussed which would incorporate other variables such as emotional factors, educational level of parents, age, and gender.

Alkhasawneh and Hargraves (2014) developed a model composed by two studies a qualitative one and a quantitative one. In each study, the factors that impact retention rates were identified, then, the critical factors were incorporated into a NN model for prediction of first year retention rates for students in science, technology, engineering, and mathematics (STEM) disciplines. The first study was a quantitative model created with the purpose of selecting the variables that had greater impact in student retention. The dataset used was comprised of 1996 student registers partitioned into two cohorts: 1468 registers of the majority of students and 498 representing data form minority groups. The genetic algorithm was used to select the variables with more impact on retention for each cohort and in this way optimize the learning time and avoid redundancy when feeding the final model (the NN). The second study was qualitative, in this part the data was collected from a focus group through an eight questions survey. In this part, content analysis was used as it is a methodology mostly applied to textual content. The results from the two studies were incorporated into a NN which was run separately to predict GPA and classify students into retained or not... The results from the NN showed an overall classification accuracy of 74%, 79% and 60% when using databases with all students, majority of students and under-represented students. Also, in was found that filtering the number of variables for each database in the quantitative model improved the classification accuracy. The research concluded that the following factors were useful for predicting performance and retention: first Math course grade, high school rank, impact of re-college intervention programs, and SAT math score.

Raju & Schumacker (2015) studied the factors of retention that lead to graduation using machine learning techniques such as LR, DT specifically C4.5, and NN. Two datasets were studied, one with precollege factors to create a prediction for completion before starting college and the second one with data collected at the end of the first semester. The model with the highest performance was logistic regression with 68.2% of classification accuracy. They also determined that the factors that have higher impact in the prediction were first semester GPA, status (full/part time), earned hours and high school GPA. Once the factors were identified the checked on the correlations with graduation rates to understand the direct impact of the factor on graduation. A ensemble of four machine learning techniques DT specifically classification and regression trees (C&RT), NN, LR, and SVM was proposed by Oztekin (2016) for the prediction of undergraduate degree completion at a four-year university. To build the model, the data was split into training and testing subsets using tenfold cross-validation, meaning that the training set was randomly divided into 10 parts, nine for training and the last for testing, this process was repeated 10 times. The model results were evaluated with overall accuracy or the percentage of correct classifications, sensitivity (recall) which is the proportion of class one correctly identified and specificity that measures the proportion of class two correctly classified. The three methods were effective in predicting degree completion, with rates over 70% for classification accuracy. The model with more consistent classification accuracy metrics was SVM. Finally, to identify the order of importance of the factors influencing degree completion within six years a fusion-based sensitivity analysis was conducted were the MSE of each model was tested with the absence of each factor. When the MSE increased significantly it meant the absent factor was of great importance. After the ranking of factors for each model was obtained, a fusion (weighted average of the ranking of all models) helped determining the final level of importance of each factor for the ensemble. The most important factors for this specific case were GPA, housing status, and the high school the student attended. The least important were ethnicity, employment status and if the student applied for financial aid.

Dissanayake et al. (2016) proposed a comparison of models for predicting student retention at St. Cloud State University. After data cleaning, the dataset for this study contained 70 variables. Principal component analysis (PCA) was used to select the variables that were not correlated with one another. Then, with the unfiltered database and the database resulting from the PCA, the study applied six prediction models: k-nearest neighbor (KNN), DT, RF, LR, NN, and BBN. The measures to evaluate the models were: overall accuracy, sensitivity, specificity, precision or percentage of correct classifications in class one from correct predictions and negative predictive value which is the percentage of correct classifications in class two from correct predictions. The results showed the models yielded better results when using the database resulting from the PCA. For instance, the RF technique presented improvement in all evaluation factors and together with LR had the highest accuracy results of 84.77% and 83.07%, respectively.

Sweeney et al. (2016) considered the importance of predicting students' grades in the courses they will enroll in during the next semester. With this purpose, they used historical transcripts and additional information from students, instructors, and courses. The methodology employed factorization machines (FM) which can be seen as an adaptation of second order polynomial regression, along with other regression techniques such as RF, stochastic gradient Descent regression (SGD), KNN, and personalized multiple linear regression (PMLP), personalized indicated that the model was used with the information of each student or course. The dataset was collected during five years from George Mason University, with a total of 15 terms including summer terms. For processing, the data was classified as transfer and non-transfer students. The factors determined to be of importance for prediction of each group were different. Further, the predictions for cold start students (first semester registered) had larger error rates. Finally, the model results indicate that MLP had the lowest error from the individual techniques; however, swapping out RF for FM when there was a lack of prior student information (cold start students) provided more accurate predictions.

Another case study using machine learning techniques was presented by Márquez-Vera et al. (2016) To predict student dropout, the authors created an algorithm called ICMR2 based on grammar based genetic programming (GBGP) where a context free grammar defines the production of the rules for classification. The new algorithm defines shorter and more accurate classification rules than the GBGP as proven by Cano et al. (2013) and it was adapted to be used with imbalanced data classes. Further, they compared the ICMR2 algorithm performance with other classification techniques as Naïve Bayes, SVM, KNN, DT C4.5. Several experiments were conducted to predict dropout in different points in time of the semester (stages zero to six). More information was available at each stage, meaning more variables were included in the prediction. Three scenarios were tested, one with all available data, another applying feature selection, and another were data resampling was allowed. The data included 419 high school students in the Academic Unit Preparatoria at university of Zacatecas Mexico. The results confirmed that as more variables were available to feed the models higher accuracy was achieved in general. In conclusion, the proposed method ICMR2 outperformed the other traditional classification algorithms. The model was able to predict dropout as early as four weeks with the highest accuracy of 83.22% and 99.8% in week 14. A set of 10 attributes provided the best performance when applying the models, which was also supported by a decrease in computational speed without risking accuracy.

In a similar study, Babić (2017) developed a classification model for predicting student academic motivation in relation to student use of the LMS. The motivation in the institution was sassed using the academic motivation scale in its college version and according to the calculated motivational average in the institution two classes were determined for the prediction: above average and below average. The methodology included the application of machine learning classifiers such as NN, DT specifically C&RT, and SVM. A test of significance applied to the classification accuracy found no evidence of a difference between the results obtained using the three methods. Therefore, their efficiency was evaluated based on their sensitivity, specificity, precision and true negative value. The research found that NN was the most efficient method to predict below-average academic motivation by predicting correctly all the examples (100% sensitivity). The study was conducted using a database comprised of information from student LMS access and student ranks on the academic motivation scale from 129 students in one year.

A comparison of methods was conducted by Tsao et al. (2017) to identify key factors that improve the accuracy of an early-alert system using different functionalities of the LMS. The data used in this study contained information on 224 students from three classes during the fall semester of 2016. The methods used for the comparison where a heuristic model and a DT. The first consisted in selecting and ranking intervals of the attributes for grouping levels (four groups 25%, 50%, 75% of students) and then, compute every combination of attribute level obtaining measures of precision and sensitivity for each. The models were created using four variables, which included average score of an online quiz, count of the course forum usage, count of roll call, and count of viewing online materials. Four different datasets where established, one for each grouping level. The study found that the differences in the results of the models in terms of precision, sensitivity, and classification accuracy were due to the different strategies of LMS use from professors. Therefore, the variables used from the LMS greatly impact the performance of the prediction models.

Pereira and Zambrano (2017) proposed a model using DT to identify patterns of undergraduate student dropout in different programs from the University of Nariño in Pasto, Colombia. The model used 6,870 student records collected from 2004 to 2011. After the data cleaning process, 31 relevant attributes were selected and classified into socioeconomic or academic factors. The results of the study identified that the most relevant academic factors were GPA, number of failed classes, department of studies, and campus location. While the relevant socioeconomic factors were tuition, home city, marital status, and living with parents.

Machine learning techniques were employed by Kondo et al. (2017) to predict at-risk students. The dataset used was obtained from the LMS during the first semester of 2015, which was comprised of records from 202 students. The methodology consisted of using LR, SVM, and RF to predict GPA. Classes for prediction were defined as s 1 if their GPA was greater than the average minus one standard deviation and 0 otherwise, meaning the student was at risk. The models were evaluated by their precision, sensitivity, and f-measure or harmonic mean between precision and sensitivity (F-measure is equal to two times precision multiplied by sensitivity divided by their addition). Also, there was an analysis of the weekly change of the comparative importance of explanatory variables. Prediction from RF showed more stable behavior in terms of precision and sensitivity. With the weekly analysis, the model was able to identify a ranking of important variables depending of the point in time (number of weeks after the semester started) that was analyzed.

Uddin and Lee (2017) developed a model to predict a good fit in major for students to decrease dropout risk. The research was developed in three stages using academic data and data from social networks. In the first stage the authors used Pearson correlation to categorize the student into one of five groups of talent traits. Then, a second algorithm was applied to find the match with the academic program for the student. It predicts the retention rate for the student by correlating the relevant talent with the degree program. At the final stage, the algorithms were integrated into the final model called the master algorithm to quantify to quantify the target variables so it can be used to predict good fit. In this stage, Machine learning techniques such as LR, MLR, BBN and DT specifically C&RT were used. For model evaluation the authors used overall accuracy and error measures, underfitting/overfitting check and proposed a new technique to assess overall accuracy they named PERFE-ciency. This measure was created to find the net/average overall performance of the master model which was composed of several methods. The results indicated that as the data size increase the more accurate is the prediction. The proposed ensemble outperformed some well-known algorithms. Academic data used in this study was collected from students in 17 universities around the world for 8-10 years, also from an online survey, and social networks.

Miranda and Guzmán (2017) aimed to identify the reasons that determine student dropout by applying different machine learning techniques including BBN, DT, and NN. The data used in this research was provided by the Catholic University of the North for 2000 to 2013. After the cleaning process the dataset contained information on 89,056 students and 11 variables. The results showed there was no significant difference within the performance of each methodology. It was found that socioeconomic factors, such as scholarships and student loans, greatly impact retention. In addition, the factor that best explained student dropout was the results of the university selection test, which is equivalent to the SAT in the US.

Iam-On & Boongoen (2017) in their research developed new algorithms for feature selection using clusters which were called WCT and WTQ. They compared the new model to other algorithms for factor selection for example PCA, kernel PCA and other three. Two datasets were studied, before and after first year. For the prediction models they compared classification accuracy from DT specifically C4.5, Naïve Bayes, KNN and NN. The classification performance was also indicative of how well the algorithm for feature selection performed. The model with higher classification accuracy was the Neural Network (77.7%) using WCT for the database collected at the end of the first year. Another comparison study between standing alone and ensemble machine learning techniques was presented by Adejo and Connolly (2018). The purpose of the research was to identify a set of variables that accurately predict student performance. Also, it explores the potential of using ensemble techniques for the same purpose. The research data was obtained from 141 students in the University of West Scotland using three sources of information: student record system, LMS, and survey. The methodology compared the classification accuracy of models used to predict student performance: DT, NN, SVM, and ensemble. PCA was applied to identify the variables that should be used in the model. Seven models were created using different combinations of variables from different information sources. The ensemble technique using variables from the three sources showed the best accuracy at approximately 80%.

4. PRINCIPAL FINDINGS

The literature refers to the rate of students in risk of discontinuing their education as: student at risk of dropout or dropout risk, attrition risk, and completion risk. Other measures related to retention have been also used in the prediction models as for example GPA. The application of machine learning techniques to predict retention in education has been increasing in the last years. The search engines used in this review gave results of early application dated to 2010. However, it is known that earlier application of such algorithms in education are dated in 1994 with studies that compared classic statistical models with machine learning models like LR, NN, among others. These studies are not included in this review to maintain consistency with the systematic search. Figure 2 presents the yearly trend of publications found. From January to August 31, 2018 only one publication was found, as it is not representative of the entire year it was not included in the figure.

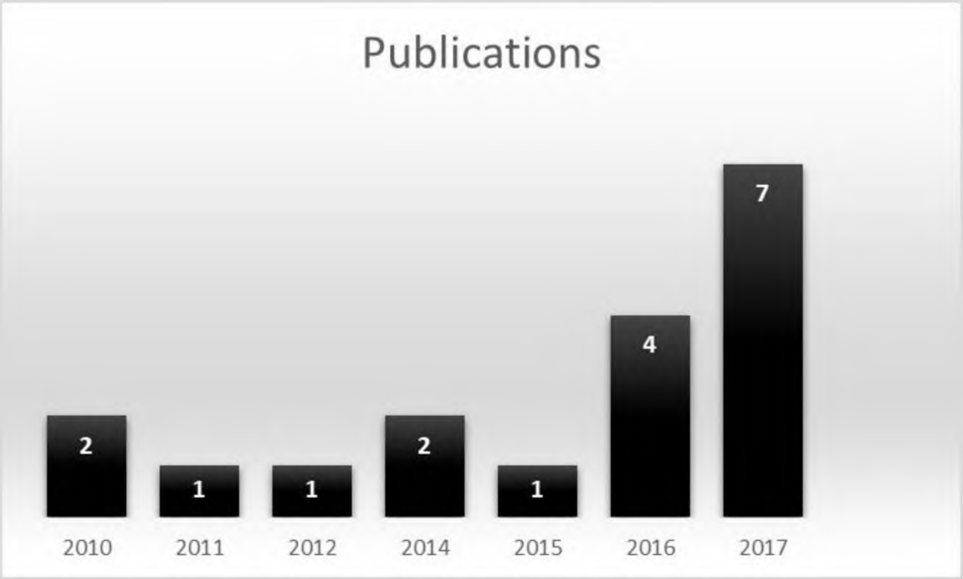


Figure 2. Yearly trend of publications about machine learning techniques applied in student retention.

A summary from literature of the machine learning techniques applied to predict retention and/or identify the main factors that impact student retention is presented in Table 1. A total review of 19 different machine learning techniques were identified in the literature. The table also presents the overall accuracy reported by the authors, specifically for datasets that in each study had a better performance.

The most frequently used techniques were NN, DT specifically C&RT, LR, SVM as presented in Figure 3. The classification accuracy ranges for the models were 71.59% - 94% for NN, 65.38% - 81.36% for DT(C&RT), 50.18% - 83.07% for LR and 57.69% - 86.4% for SVM. More consistent results were attributed to DT with a narrower range, suggesting it is a good algorithm to be applied to the topic in study, however it is not clear which method can be considered the best in general. Also, it is important to highlight that in studies that compared ensemble techniques with stand-alone techniques, such as Delen (2010) Essa and Ayad (2012), Dissanayake et al. (2016) and Sweeney et al (2016) the results were more consistent from ensembles with classification accuracy ranging between 79.36% and 81.67%, one of the narrower ranges found. This indicates that ensembles could be more efficient methods to predict student dropout risk. Nevertheless, there were not a broad number of papers to determine which kind of ensemble has better performance.

Table 1. Machine Learning techniques employed for prediction of student retention

Method	**Study**	**Model performance (Overall accuracy)**
Bayesian Belief Network	Slim et al. (2014)	MSE curves
	Dissanayake et al. (2016)	85.27%
	Miranda and Guzmán (2017)	76%
	Uddin and Lee (2017)	Accuracy itself was not reported.
Boosted trees (Ensemble-boosting)	Delen (2010)	80.21%
Decision tree (CHAID)	Raju & Schumacker (2015)	73.50%
Decision tree (C&RT)	Oztekin (2016)	73.75%
	Dissanayake et al. (2016)	81.36%
	Babić (2017)	65.38%
	Tsao et al. (2017)	68.25%
	Pereira et al. (2017)	80%
	Uddin and Lee (2017)	Accuracy itself was not reported.
	Miranda and Guzmán (2017)	74%
	Adejo and Connolly (2018)	78%

Table 1. Machine Learning techniques employed for prediction of student retention (Cont.)

Decision tree (C4.5)	Márquez-Vera et al. (2016)	86.40%
	Iam-On & Boongoen (2017)	92.60%
Decision tree (C5)	Delen (2010)	80.65%
	Delen (2011)	78.25%
Factorization machine	Sweeney et al (2016)	74.23%
ICMR2	Márquez-Vera et al. (2016)	78.20%
Information fusion (Ensemble stacking)	Delen (2010)	82.10%
K-Nearest neighbor	Dissanayake et al. (2016)	83.37%
	Sweeney et al (2016)	80.61%
	Márquez-Vera et al. (2016)	84.20%
	Iam-On & Boongoen (2017)	93.60%
Logistic regression	Delen (2010)	74.26%
	Delen (2011)	74.33%
	Raju & Schumacker (2015)	77.10%
	Oztekin (2016)	50.18%
	Dissanayake et al. (2016)	83.07%
	Kondo et al. (2017)	75%
	Uddin and Lee (2017)	Accuracy itself was not reported.
Naïve Bayes	Márquez-Vera et al. (2016)	78.30%
	Iam-On & Boongoen (2017)	93.80%
Neural networks	Delen (2010)	79.85%
	Delen (2011)	81.19%
	Alkhasawneh and Hargraves (2014)	79.00%
	Raju & Schumacker (2015)	77.70%
	Oztekin (2016)	71.59%
	Dissanayake et al. (2016)	84.87%
	Babić (2017)	76.92%
	Miranda and Guzmán (2017)	83%
	Iam-On & Boongoen (2017)	94%
	Adejo and Connolly (2018)	73%
Multiple linear regression	Sweeney et al (2016)	78.86%
	Uddin and Lee (2017)	Accuracy itself was not reported.
Random forest (Ensemble-bagging)	Delen (2010)	81.80%
	Dissanayake et al. (2016)	85.87%
	Sweeney et al (2016)	79.36%

Table 1. Machine Learning techniques employed for prediction of student retention (Cont.)

Stochastic Gradient Descend	Sweeney et al (2016)	82.07%
Simulation	Slim et al. (2014)	MSE curves
Support vector machines	McAleer and Szakas (2010)	79.59%
	Delen (2010)	81.18%
	Oztekin (2016)	77.61%
	Márquez-Vera et al. (2016)	86.40%
	Babić (2017)	57.69%
	Kondo et al. (2017)	65%
	Adejo and Connolly (2018)	83%
SVM+DT+NN (Ensemble stacking)	Adejo and Connolly (2018)	81.67%

Consistency could seem a good indicator to determine the better methodology , but recalling Section 3 of this paper, the information used in the prediction models varies depending on the goal of the study, for instance, Slim et al. (2014), Márquez-Vera et al. (2016), and Tsao et al. (2017) wanted prediction results early in the career by week, by semester or even by year (varying by study). While Essa and Ayad (2012), Sweeney et al. (2016) and Tsao et al. (2017) where predicting risk of dropout for different courses using factors not only related to the student but also to the courses. Even when these studies shared the goal, the set of variables changes. Thus, it would not be appropriate to indicate there is a better machine learning technique to be applied for student retention from the information found in this systematic review. However, it can be concluded that machine learning techniques, in general, offer good classification accuracy with an average of 78% in a range between 50.18% and 94%.

Determining the factors that most influence degree completion was a common objective in the different studies in this systematic review. Table 2 presents a summary of the factors that showed high impact on prediction of student retention in the different studies. The list was organized by categories as different names attempting a common variable were used in the different studies. Also, the frequency in which the variable was used was presented (No. of references in the table).

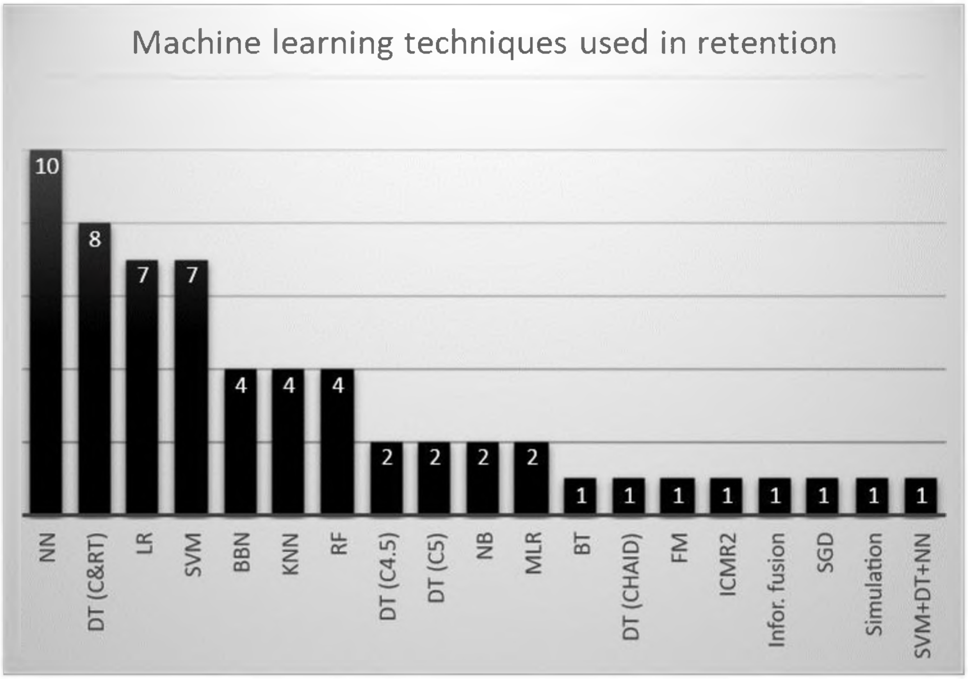


Figure 3. Frequency of use of Machine learning techniques to predict student retention

Table 2. Factors with high impact on prediction of retention in the literature

Category/ factor	Reference	No. of references
College GPA		
Fall GPA	Delen (2010), Delen (2011), Oztekin (2016)	3
Overall GPA	Slim et al. (2014), Dissanayake et al. (2016), Pereira & Zambrano (2017), Miranda & Guzmán (2017)	3
GPA 100 Level classes	McAleer & Szakas (2010)	1
GPA 200 Level classes	McAleer & Szakas (2010)	1

Table 2. Factors with high impact on prediction of retention in the literature (Cont.)

First semester GPA	Raju & Schumacker (2015)	1
Spring GPA	Oztekin (2016)	1
Previous term GPA	Dissanayake et al. (2016)	1
Aggregate GPA for total students enrolled until previous term	Dissanayake et al. (2016)	1
Aggregate GPA for total students enrolled since first offered	Dissanayake et al. (2016)	1
Before starting college		
High school GPA	Raju & Schumacker (2015), Dissanayake et al. (2016), Márquez-Vera et al. (2016)	3
High school attended	Alkhasawneh & Hargraves (2014), Oztekin (2016)	2
Mothers level of education	Márquez-Vera et al. (2016)	1
living with parents	Pereira & Zambrano (2017)	1
Home city	Pereira & Zambrano (2017)	1
Major preference before admission	Miranda & Guzmán (2017)	1
Financial aid		
Fall student loan	Delen (2010), Delen (2011)	2
Spring student loan	Delen (2010), Delen (2011)	2
Spring grant/tuition waiver/scholarship	Delen (2010), Delen (2011)	2
Student benefits	Miranda & Guzmán (2017)	1
SAT		
SAT comprehensive	Delen (2011), Miranda & Guzmán (2017)	2
SAT math	Alkhasawneh & Hargraves (2014), Márquez-Vera et al. (2016)	2
Number of credits		
Earned by registered (EarnedHrs/RegisteredHrs)	Delen (2010), Delen (2011)	2
Fall Hours registered	Delen (2010)	1

Table 2. Factors with high impact on prediction of retention in the literature (Cont.)

Credits earned at the end of 1st semester	Raju & Schumacker (2015)	1
Spring Hours registered	Oztekin (2016)	1
Credits enrolled current term	Dissanayake et al. (2016)	1
Total credits earned	Dissanayake et al. (2016)	1
Total credit hours attempted	Dissanayake et al. (2016)	1
Collected from LSM		
Assignment view	Babić (2017)	1
Forum view discussion	Babić (2017)	1
Questionnaire view	Babić (2017)	1
Resource view	Babić (2017)	1
Duration of logging-in time	Kondo et al. (2017)	1
Collected from surveys		
Impact of pre-college intervention programs	Alkhasawneh & Hargraves (2014)	1
Level of motivation	Márquez-Vera et al. (2016)	1
Preferred place for studying	Márquez-Vera et al. (2016)	1
regular consumption of alcohol	Márquez-Vera et al. (2016)	1
smoking habits	Márquez-Vera et al. (2016)	1
Having an administrative sanction	Márquez-Vera et al. (2016)	1
Other factors		
Marital status	Delen (2010), Pereira & Zambrano (2017)	2
Work	Raju & Schumacker (2015), Márquez-Vera et al. (2016)	2
Housing status	Raju & Schumacker (2015), Oztekin (2016)	2
Attendance	Márquez-Vera et al. (2016), Kondo et al. (2017)	2
College	Raju & Schumacker (2015), Oztekin (2016)	2
Gender	Raju & Schumacker (2015)	1

Table 2. Factors with high impact on prediction of retention in the literature (Cont.)

Status- Full or part time	Raju & Schumacker (2015)	1
Zip code	Dissanayake et al. (2016)	1
Age	Márquez-Vera et al. (2016)	1
Tuition	Pereira & Zambrano (2017)	1
Campus location	Pereira & Zambrano (2017)	1
Department	Pereira & Zambrano (2017)	1
Failed courses	Pereira & Zambrano (2017)	1
First Math. course grade	Alkhasawneh & Hargraves (2014)	1
Fall completion rate per semester	Oztekin (2016)	1
Course related studies		
Instructor role type (Adjunct, FT, PT, GRA, GTA)	Dissanayake et al. (2016)	1
Course ID	Márquez-Vera et al. (2016)	1
Num. students enrolled in the course current term	Márquez-Vera et al. (2016)	1

Literature indicates that the importance of factors changed according to the institution and the methodology applied. One of the reasons behind such differences is that the studies have different goals like predicting for specific course, for a different period (as mentioned before), implementing a new algorithm for variable selection Iam-On & Boongoen (2017), this among others. Then, the sets of factors used in each study was different. The result of this, is that most of the factors were used in no more than one study, maximum in three studies very few of them. In other words, from the information in this systematic review it cannot be determined a set of variables that can be generalized and applied universally to any institution for retention prediction. Meaning, the results of each model depend on the information available for the study, specifically when referring to the classification of importance of the variables used.

However, it was possible to identify different categories of information that showed to be important across all the studies included in this review (refer to Table 2). In fact, it is evidenced that GPA is of great importance for student retention identified important in 63% of the studies, followed by "before starting college" factors (i.e. high school GPA) found important in 47% of the studies, and financial aid found important in 37% of the studies.

It was found in the literature that there is a considerable interest in creating models to predict student risk of dropout early in the student career such that retention strategies can be more effective if the student is identified at risk as early as possible. For example, some researchers, such as Kondo et al. (2017) and Márquez-Vera et al. (2016), even studied prediction on a weekly basis since the highest dropout rates occur during the first year of college.

In addition, a common statement from researchers was the use of the results should be a guide to create strategies focused on individual needs (Delen (2010), Raju & Schumacker (2015), Márquez-Vera et al. (2016); Essa and Ayad, (2012)), . For example, Essa and Ayad (2012) considered early detection of dropout risk important and generated strategies that focused on the combination of important factors for student dropout risk. Their proposed model was a tracking system of the individuals and the retention strategies applied to each student.

Some institutions have already benefitted from the use of machine learning techniques in the identification of students at risk of dropout and the results show important increases in retention rates. However, the development of such prediction models requires an enormous effort in the administration of the data collection and analysis. For example, in the model developed for Georgia State University, 800 variables were employed to identify student performance. From the amount of correlations created, the causes and weaknesses that prevent student from having satisfactory performance were targeted. For instance, a low score in math in high school will have and important negative impact on student risk of dropout in the early stages of the studies. Thus, by identifying this correlation the institution could create intentional strategies such as giving a conditional enrollment that requires the student to be tutored in math in the first semester, or even before starting (McMurtrie, 2018; Dimeo, 2017).

5. CONCLUSIONS AND FUTURE WORK

Machine learning techniques have been applied in education to predict retention and identify factors that influence retention rates for several years, with more successful results since 2010. The research area is relatively new and is still a work in progress. More research is needed to determine the factors that impact student retention and to define and architect systems that allow for educational institutions to be alerted when to implement retention strategies and what strategies are most appropriate for each student.

The advantages of data collection offered specifically by an LMS in institutions has been and should continue to be a main source of information. As presented in the literature, important factors that influence the identification of students at risk were drawn from that source. Further, LMS provide up-to-date information, which is an opportunity to create models that can provide timely feedback and notification.

The most frequently used techniques were DT, NN, and SVM with performance rates over 67%. Also, other models such as ensembles have been developed that have shown accurate classifications (80% and higher). However, only a few studies use ensembles and it is not conclusive that they represent a better option for the prediction of student retention. Future research should focus on using ensemble techniques to nurture the body of knowledge on what mixtures of machine learning techniques can provide higher accuracy.

It was also found that although novelty models have been developed, they must be customized for each institution. The ranking of factors in the models change depending on the list of factors selected for the study. A list of factors that can be universally applied for prediction of degree completion has not been identified in the literature.

Institutions should develop synchronized systems that are able to collect student data that feed the learning algorithms in order to have the most benefit from them. As it is statistically assumed, the more data the more reliable are the results. However, it is also important to highlight from this systematic review that the algorithms have proved to be efficient for predicting student success using less than 68 variables. This means that the studies can be segmented, and specific datasets can lead to specific analysis. As stated by Essa and Ayad (2012) “Decomposition provides a flexible mechanism for building predictive models for application in multiple contexts” Meaning bey decomposition the application of the model in different scenarios of the institutions.

This systematic review contributed with the analysis of the existing literature took from the specific search engines mentioned in the methodology section of this paper. However, it is limited to the time frame also specified in the same section, and to the search engines available in the Missouri University of Science and Technology search engine portfolio. Therefore, as future work it is recommended to include literature produced after August 31, 2018 together with studies from additional search engines.

REFERENCES

Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education, 10*(1), 61-75.

Alkhasawneh, R., & Hargraves, R. H. (2014). Developing a hybrid model to predict student first year retention in STEM disciplines using machine learning techniques. *Journal of STEM Education: Innovations and Research*, *15*(3), 35-42.

Babić, I. D. (2017). Machine learning methods in predicting the student academic motivation. *Croatian Operational Research Review*. *8*, 443-461.

Cano, A., Zafra, A., & Ventura, S. (2013). An interpretable classification rule mining algorithm. *Information Sciences*, *240*, 1-20.

Cardona, T., Cudney, E., & Snyder, J. (2019a). *Predicting degree completion through data mining*. Proceedings of the ASEE Annual Conference & Exposition, Tampa, FL.

Cardona, T., Cudney, E., & Snyder, J. (2019b). *Predicting student retention using artificial neural networks*. Proceedings of the IISE Annual Conference, Orlando, FL.

Cardona, T., & Cudney, E. (2019). *Predicting student retention using support vector machines*. Proceedings of the International Conference on Production Research Manufacturing Innovation: Cyber Physical Manufacturing, Chicago, IL.

Carnevale, A. P., Jayasundera, T., & Gulish, A. (2016). America's Divided Recovery: College Haves and Have-Nots. Georgetown University Center on Education and the Workforce. https://cew.georgetown.edu/wp-content/uploads/Americas-Divided-Recovery-web.pdf

Chen, X. (2013). STEM attrition: College students' paths into and out of STEM fields. Statistical Analysis Report. NCES 2014-001. *National Center for Education Statistics*.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498-506.

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice*, *13*(1), 17-35.

Dimeo, J. (2017). Georgia State improves student outcomes with data. *Inside Higher Education*. Retrieved from https://www.insidehighered.com/digital-learning/article/2017/07/19/georgia-state-improves-student-outcomes-data

Dissanayake, H., Robinson, D., & Al-Azzam, O. (2016, January). Predictive Modeling for Student Retention at St. Cloud State University. In *Proceedings of the International Conference on Data Mining (DMIN)* (p. 215). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Essa, A., & Ayad, H. (2012). Improving student success using predictive models and data visualisations. *Research in Learning Technology*. *20*, 58-70.

Iam-On, N., & Boongoen, T. (2017). Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings. *International Journal of Machine Learning and Cybernetics*, *8*(2), 497-510.

Hoffman, E., Starobin, S., Laanan, F. S., & Rivera, M. (2010). Role of community colleges in STEM education: Thoughts on implications for policy, practice, and future research. *Journal of Women and Minorities in Science and Engineering*. *16*(1), 85-96.

Kirp, D. (2019). The college dropout scandal. The Chronicle Review. https://www.chronicle.com/interactives/20190726-dropout-scandal

Kondo, N., Okubo, M., & Hatanaka, T. (2017, July). Early detection of at-risk students using machine learning based on LMS log data. In *Advanced Applied Informatics (IIAI-AAI), 2017 6th IIAI International Congress on* (pp. 198-201). IEEE.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, *33*(1), 107-124.

McAleer, B., & Szakas, J. S. (2010). Myth Busting: Using Data Mining to Refute Link between Transfer Students and Retention Risk. *Information Systems Education Journal*, *8*(19), 3-7.

McMurtrie, B. (2018). Georgia State U. made its graduation rate jump. *The Chronicle of Higher Education. Retrieved from https://www. chronicle. com/article/Georgia-State-U-Made-Its/243514.*

Miranda, M. A., & Guzmán, J. (2017). Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos. *Formación universitaria*, *10*(3), 61-68.

Morris, L. V. (2016). Mining Data for Student Success. *Innovative Higher Education*, *41*(3), 183-185.

National Center for Education Statistics (2019). Undergraduate retention and graduation rates. https://nces.ed.gov/programs/coe/indicator_ctr.asp

National Science Board. (2016). *Science and engineering indicators 2016*. Arlington, VA. National Science Foundation.

Oztekin, A. (2016). A hybrid data analytic approach to predict college graduation status and its determinative factors. *Industrial Management & Data Systems*, *116*(8), 1678-1699.

Pereira, R. T., & Zambrano, J. C. (2017, December). Application of decision trees for detection of student dropout profiles. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on* (pp. 528-531). IEEE.

Raju, D., & Schumacker, R. (2015). Exploring student characteristics of retention that lead to graduation in higher education using data mining models. *Journal of College Student Retention: Research, Theory & Practice*, *16*(4), 563-591.

Restuccia, D., & Taska, B. (2018). Different skills, different gaps: Measuring and closing the skills gap. Developing Skills in a Changing World of Work: Concepts, Measurement and Data Applied in Regional and Local Labour Market Monitoring Across Europe, 207. https://www.burning-glass.com/research-project/skills-gap-different-skills-different-gaps/

Shapiro, D., Dundar, A., Huie, F., Wakhungu, P. K., Bhimdiwala, A., & Wilson, S. E. (2018). Completing college: A national view of student completion rates – Fall 2012 cohort (Signature Report 16). Herndon, VA: National Student Clearinghouse Research Center.

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

Snyder, J., & Cudney, E. (2017). Retention models for STEM majors and alignment to community colleges: A review of the literature. *Journal of STEM Education*, *18*(3), 30-39.

Snyder, J., and Cudney, E. (2018). *A retention model for community college STEM students*. Proceedings of the ASEE Annual Conference & Exposition, Salt Lake City, UT.

Sweeney, M., Lester, J., Rangwala, H., & Johri, A. (2016). Next-Term Student Performance Prediction: A Recommender Systems Approach. *JEDM| Journal of Educational Data Mining*, *8*(1), 22-51.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British journal of management*, *14*(3), 207-222.

Tsao, N. L., Kuo, C. H., Guo, T. L., & Sun, T. J. (2017, July). Data Consideration for At-Risk Students Early Alert. In *Advanced Applied Informatics (IIAI-AAI), 2017 6th IIAI International Congress on* (pp. 208-211). IEEE.

Uddin, M. F., & Lee, J. (2017). Proposing stochastic probability-based math model and algorithms utilizing social networking and academic data for good fit students prediction. *Social Network Analysis and Mining*, *7*(1), 29.

Williford, A. M., & Schaller, J. Y. (2005, May). All retention all the time: How institutional research can synthesize information and influence retention practices. In *Proceedings of the 45th Annual Forum of the Association for Institutional Research.*

II. HIGHER EDUCATION STUDENT SUCCESS: A SYSTEM TO EVALUATE DEGREE COMPLETION

Tatiana A. Cardona[a], Elizabeth A. Cudney[a], Jennifer Snyder[b], and Roger W. Hoerl[c]

[a]Deparment of Engineering Management and Systems Engineering, Missouri University of Science and Technology

[b]School of Science, Valencia College's East campus.

[c]Union College, Schenectady, NY.

ABSTRACT

The goals of higher education have evolved thought time based on the impact that technology development and industry have on productivity. Nowadays, jobs demanded increased technical skills, and the supply of prepared personal to assume those jobs was insufficient. The system of higher education needs to evaluate their practices to realize the potential of cultivating an educated and technically skilled workforce. Currently, completion rates are for universities are very low to accomplish the aim of closing the workforce gap. Only 40% of freshmen will graduate. And, for community college graduation rates are even lower. The reasoning behind these statistics is multifaceted and will not be solved without concerted effort from all partners in higher education. In recent years machine learning techniques have been applied to analyze student data, which aligns with the focus of improving the processing of information through data mining. The primary objective of this research is to stablish the agents that intervene in student success, not as separate matters but as an integrated system that allows for the application of machine learning for student success prediction. In addition, the proposed system and

a mix of the agents' interactions was evaluated to determine the accuracy of student success predictions using neural networks (NN) technique.

1. INTRODUCTION

The goals of access in higher education have evolved through time. Further, they have changed based on the impact that technology development and industry have on U.S. productivity [1]. Around the mid-20th century important changes were made in higher education to improve enrollment and equal access for all socioeconomic classes [2]. However, during the turn of the century, when the universities began increasing their capacities to handle higher enrollment rates, another situation emerged; students were taking longer than expected to graduate or did not graduate at all [3]. Research started reporting a gap in the workforce, jobs demanded increased technical skills, and the supply of prepared personal to assume those jobs was insufficient [4]. The government, universities, and community colleges started to turn their attention towards student success and, consequently, student completion rates and retention [5].

Low completion rates gained considerable attention in scientific research where they started to be studied and addressed as student persistence [6], [7]. For example, Tinto [6]-[8] presented student persistence in three dimensions: commitment to the institution, academic goals, and career goals. Today, these are still considered the basis student success approaches. The extent to which these dimensions have been studied has become expansive and additional factors involved in student success have been identified. And some attempts have been made to generate solutions to low rates of student success [9]. Nevertheless, higher education institutions still struggle with low completion and the workforce gap is widening at faster rates [1], [10]. The National Student Clearinghouse Research Center [11] indicates that 65.7% of students at four-year public institutions graduate within six years, while the number decreases dramatically for two years institution where only 39.2% of students in graduate within three years. According to Kirp [[12]], 40% of college freshman will not graduate. Further investigation of graduation data indicates that nationally only 17% of students that start at a community college will transfer and graduate from a four-year institution within six years [13]. The rise in students attending community colleges, or two-year institutions, during their first two years has resulted in a need to include this important agent of the higher education in the analysis [14].

Increasing student retention in higher education is of important interest as it is a step forward in terms of decreasing the skill gap in the workforce, and it also reflects institutional commitment to the students [14]. But what if possible solutions have been identified, implemented and the rates of completion are still low? The current literature suggests that the scope of the investigation and implementation of possible solutions has been an important aspect of failing to improve student success [17]. Institutions have concentrated their efforts on studying small segments of students such as minority, low income students, first generation college students, and freshmen, among other groups [4], [18]-[22]. Although imperative insights can be obtained from such studies, and they represent benefits for each specific segment, the results for the broader view have not been very promising [23]. The implementation of reforms and strategies should be done in a progressive manner to provide evidence of improvement of student completion [24].

Thus, the focus changed, and experts have proposed a more holistic emphasis. They suggested that the new perception of student success should be concentrated on helping students define and meet their educational goals [25], and preparing them to support themselves and achieve what they envision for their future [26].To support this perception, three major factors come into play [24], [26]-[29]. The first factor is the culture of student success. Reforms need to take place to integrate the system forces (e.g., management, administrative stuff, and faculty) to make student success a priority and intrinsic part of an educational institution's strategies and activities, rather than a secondary project. Second, information technology (IT) should be recognized as an important agent for improving student success in three key aspects. First, data collection through the synchronization of systems can provide data such as the learning management system (LMS) and enrollment system as well. Also, additional mechanisms for data collection should be implemented (e.g., wearables) or more sophisticated measures such as virtual reality (VR) for new class modalities (offered by virtual means) and artificial intelligence (AI) companions (e.g., robots). Second, the implementation of data analytics methodologies through the creation of software or programing developments that allow predictions and a flag system for students at risk of attrition will enable institutions to focus retention strategies. And third, the scalability, results must be scalable, and the applications must be able to be successfully applied in other institutions.

The development of new technologies that support machine learning techniques and AI can help achieve scalability [30]. However, for data analytics function as an important aspect to improve student success, certain things need to happen. Before starting to develop reforms and strategies to improve student success, mechanics of the system must be identified to recognize the different ways student retention can be evaluated. Also, the institution should determine the interactions that systematically impact student success. The primary objective of this research is to establish the agents that intervene in student success, not as separated aspects but as an integrated system through the application of machine learning. To accomplish this, the proposed system and a mix of the agents' interactions will be tested to determine if they can produce accurate student success predictions. Therefore, the major contributions of this research can be summarized as follows:

The formulation of a complex system structure that allows for the evaluation of strategies to improve student success in higher education specifically community college and university environments. The model will represent the information flow of a student that enters the higher education system.

Implementation of neural networks (NN) techniques using the proposed system to validate its structure to identify the level of impact of the factors selected for the model and obtain a prediction of potential students at risk of attrition.

The remainder of the paper is organized as follows. Section 2 presents the architecture of the proposed system followed by its validation using a NN model in Section 3. Finally, Section 4 presents the conclusions, limitations, and future work.

2. HIGHER EDUCATION SYSTEM TO EVALUATE STUDENT SUCCESS

The representation of a system and its flow of information into models of analysis using machine learning techniques generates important insights in the system behavior, patterns, and inherent features. This creates a basis for decision making, control, management, and transformation of the system under investigation [31], [32]. In the case of student success, a system should represent the integration of the key factors to enable students to accomplish their educational goals. This will allow for the development of strategies and implementation of reforms that are more appropriate to each institution.

Therefore, the framework for the development of reforms towards student success should be based on a system that represents the interactions of a student within an institution (in higher education) based on an institutional culture of student success, an evidence-base culture (IT structure and support), and a projection of the scalability of such reforms.

2.1. SYSTEM DESCRIPTION

The development of a system that represents the student within the institution is an important instrument for the identification of mechanics and interactions that systematically impact student success.

The proposed system is an intent to achieve this aim to establish the agents that intervene in student success. Also, to offer a clearer structure for the creation of models that allow for the evaluation and application of reforms for improving completion rates using machine learning techniques.

For the purposes of this study, student success is defined in terms of the attainment of educational objectives [34], specifically student completion of a program within a certain amount of time. The time considered was 150% of the designed time for completion. This period was defined to be consistent with the 1990 Student Right-to-

Know Act, which requires postsecondary institutions to report the rate of students graduating in 150% of the time the program was designed to be completed within [35]. For instance, a student in an associate degree program should complete the degree program in two years. However, a student is considered successful if they complete the studies in three years or less. For a bachelor's program, a successful student completes a degree is six years or less.

2.2. METHODOLOGY

First, by consulting the literature it was possible to identify the factors that intervene in student completion of higher education. Next, the flow of information was established, and the structure of the system was developed. Second, to validate the structure of the system, a model to predict student success was developed. A NN was developed using the factors established for the system.

2.3. FACTORS

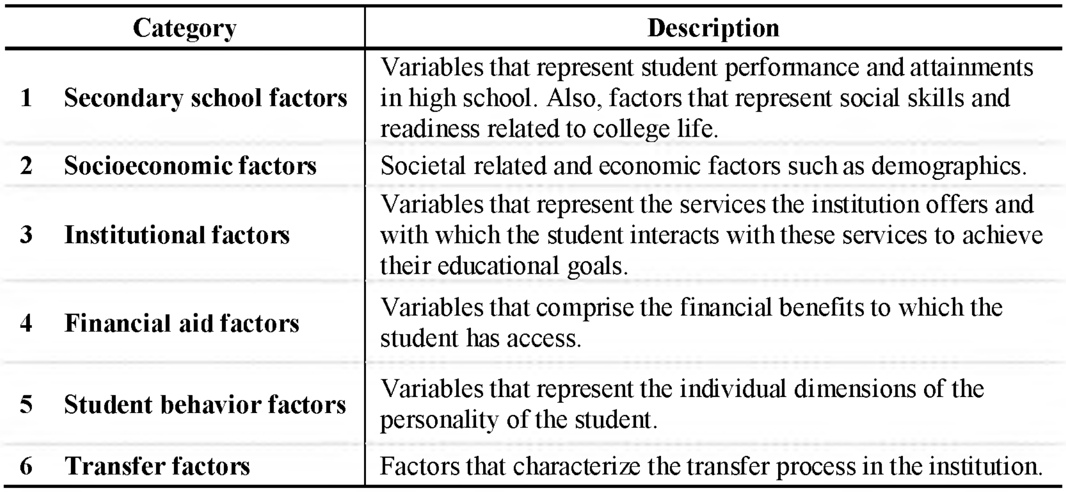
In reviewing the literature [35]-[37], it was possible to establish the factors that impact student completion. To develop the system architecture those factors were classified into six categories as shown in Table 1.

2.4. SYSTEM STRUCTURE

Figure 1 represents the architecture of the higher education system. It is comprised of several inputs that represent the status of the student before entering the higher education system, which include the secondary school and socioeconomic factors.

This information gives the institution a starting point to evaluate the potential of the student to succeed. Here, admission requirements and other policies determine the entrance of the individual to the system. Once in the institution, the interactions between the institutional, financial and/or transfer factors, and behavioral factors allow the transformation of the student characteristics to obtain an output, which is declared as degree completion.

Table 1. Categories of factors that impact in student success



	Category	Description
1	**Secondary school factors**	Variables that represent student performance and attainments in high school. Also, factors that represent social skills and readiness related to college life.
2	**Socioeconomic factors**	Societal related and economic factors such as demographics.
3	**Institutional factors**	Variables that represent the services the institution offers and with which the student interacts with these services to achieve their educational goals.
4	**Financial aid factors**	Variables that comprise the financial benefits to which the student has access.
5	**Student behavior factors**	Variables that represent the individual dimensions of the personality of the student.
6	**Transfer factors**	Factors that characterize the transfer process in the institution.

In summary, the system represents the characteristics a student possesses prior to entering the higher education system and within the system to be able to complete (or not complete) their degree.

The system also contains the IT department as a transversal agent. It represents the platform for data collection and analysis. A more detailed description of the inputs, process, and outputs of the system is presented in the following sections.

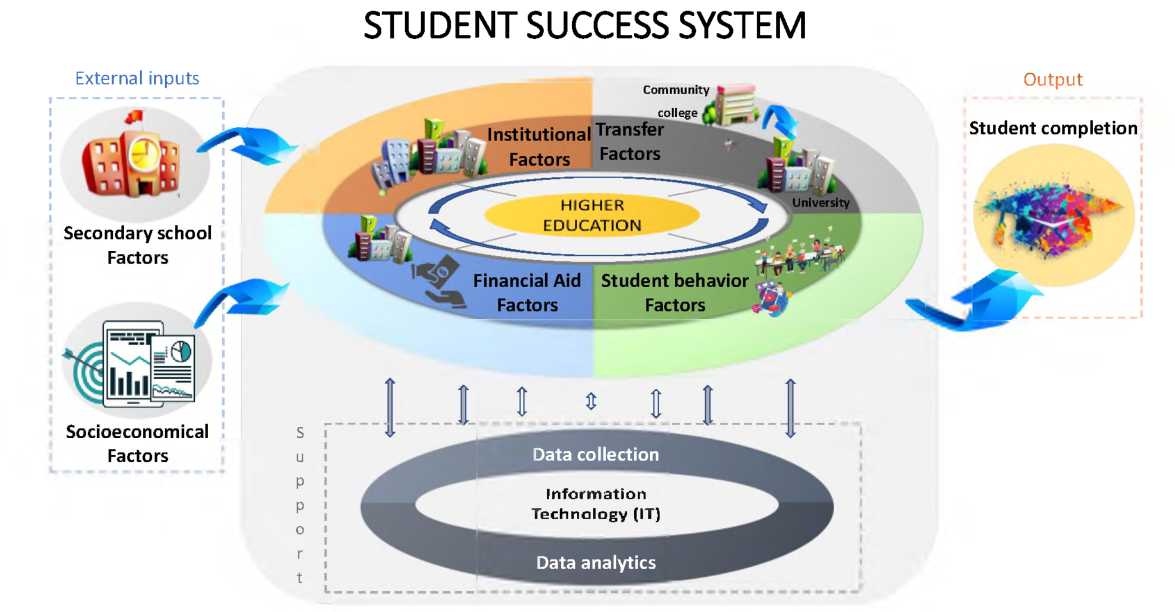


Figure 1. Higher education architecture for prediction of student success

2.4.1. Factors Interactions, System Rules. The interactions of the factors are defined by the institutional policies and rules established by the institution. For example, admission requirements and completion requirements as shown in Table 2.

Table 2. Example of system rules

Category	Policy/rule
Admission requirements	Minimum grade point average (GPA)
	Minimum score for standard entry test
	Minimum financial resources to cover at least a year of studies (e.g., tuition, boarding, alimentation, and university fees)
Completion requirements	Range of credits allowed to take in a semester (min-max)
	Minimum number of credits to graduate
	Classes required for graduation

2.4.2. External Factors or System Inputs. External factors or system inputs are usually collected during the student's admission process and most will not change through the system's interactions. For the prediction model these factors are considered static factors. The categories for secondary school factors and socioeconomic factors and examples of each are presented in Table 3 and Table 4, respectively.

Table 3. Secondary school factors

Category	**Subcategory**	**Example(s)**
Secondary school	Academic attainment	Scores on standardized higher education entry exams
		GPA from high school
		Performance awards
	College readiness	Social integration
		Motivation to learn
		Participation in outreach activities
		Participation in precollege intervention programs
		First generation to attend college
		Major preference

Table 4. Socioeconomic factors

Category	**Subcategory**	**Example(s)**
Socioeconomic	Demographics	Age
		Gender
		Economic status
		Marital status
	Family and peer support	Financial support
		Parental encouragement
		Parents level of education

2.4.3. Internal Interaction Factors. Internal interaction factors are, in their majority, in constant evolution as the result of the student interactions within the institution. Subsequently, these factors can define the holistic view of the characteristics of student success for an institution. These factors can be broken down as institutional, financial aid, student behavior, and transfer factors; which are presented with examples in Table 5 through Table 8, respectively.

Table 5. Institutional factors

Category	**Subcategory**	**Example(s)**
Institutional	Teaching	Instructor experience: time teaching, time teaching a certain course
		Instructor professional development
		Instructor's pedagogical preparation
		Instructor workload
		Instructor role type (adjunct, full time, part time, graduate research assistant, graduate teaching assistant)
	Pathway design	Curriculum and design of core courses
		Course design, course content and orientation (e.g., area of reference, pedagogical approach)
		Number of students enrolled in the course
	Peer involvement	Orientation program
		Instructor intervention for intentional academic advice and development
	Campus environment	Institutional policies
		Specific student support, in aspects different than academics (e.g. counseling, financial counseling and literacy)
	Multidimensional	Promote culture of diversity

2.4.4. Informational Technology Support. IT supports the system with the administration and maintenance of the infrastructure for data collection and data analysis platforms. The importance of this department is for it to allow the synchronization of the different informational sources such as the LMS and the different modalities established by the institution to collect data such as information (ID) card tracing and VR experience. The information collected is the basis of the development of the prediction models.

Table 6. Financial aid factors

Category	Subcategory	Example(s)
Financial aid	Student benefits	Scholarships and grants
		Waiver programs
		Awards
	Loans	Student loans
	Emergency funds	Food pantry
		Emergency funding

Table 7. Student behavior factors

Category	Subcategory	Example(s)
Student behavior	Academic attainment	GPA
		Credits enrolled in certain amount of time
		Full or part time
		Time to graduation
		Study progression (e.g., first, second, third, or fourth year)
		Failed courses and other performance measures
	Academic interaction	Variables related to the usage of the LMS (e.g., log in duration, items visited during log in)
		Participation in on-campus activities, student organizations
	Academic preparation	Study habits
		Hours of study outside the university
		Days of study before a test
		Study mode (e.g., on campus, distance)
		Preferred place for studying
		Attendance
	Student engagement	Level of motivation
		Overall satisfaction with the institution
		Willingness to attend the institution again
		Perception of institutional quality

Table 8. Transfer factors

Category	Subcategory	Example(s)
Transfer	Academic attainment	GPA
		Credits accumulated
		Failed courses and other performance requirements
	Student goals	Reasons for transferring
		Institutional alignment of college with university

It is also important to mention that the improvement of an IT structure to support the student success system has several issues that need to be addressed such as ethical issues; however, the discussion of these issues is out of the scope of this study.

3. MODEL VALIDATION

Machine learning techniques have been proven to be an adequate approach to predict student success [37]-[41]. As effective models can continuously learn from the data, these models help to determine if the student is at risk prior to the student leaving the institution. Those models surpass the survey methodologies that could serve as an instrument for detecting patterns but only at a snapshot in time.

It is important to highlight that this validation refers to the proposed architecture, and at this point is not intended for the creation of strategies to improve student completion. This due to the limitations in the information collected. However, it is possible to infer the effectiveness of the architecture by creating a prediction model. The prediction model indicates that the factors selected have an impact on the system, and the structure of the interactions are adequate for future modeling.

As previously mentioned, the focus of modeling student success should be progressive and holistic such that reforms and strategies can be developed from the results. This will enable improvements in completion rates, and the models can be scalable to other institutions. Further, once the IT platforms are at the service of the system, the information collected generate value in several ways. This is another use of the proposed system. Sub models can be developed to characterize relevant interactions in the system. The level of granularity, specification, and the segment selected for the models should be determined according to established goals; this will help avoid inadequate results or find misrepresented behaviors, unnecessary incurrence in complexity and cost and possible delays [31]-[32].

3.1. VALIDATION METHODOLOGY

To validate the proposed architecture, a trial model for classification of students was prepared using the NN technique. Factors from different categories in the system were selected according to the availability of information to create the dataset. Finally, the model was assessed using performance measures such as overall classification accuracy, precision, and recall.

NN was selected for this study as it currently is the most widely used machine learning technique for student success predictions. Also, NN has shown better performance in the classification of student success in comparison with support vector machines and decision trees [42]-[47].

3.2. DATA

Public information for a bachelor's degree from a university in the Midwest was selected. Statistical reports and published studies of the institution in fall 2017 were used to create a database of 10,000 entries. The rules or interactions within the factors were defined based on institutional policies. From the six categories established for the proposed system, it was possible to characterize factors in secondary school, student behavior, and financial aid categories. Detailed information about the variables selected for the model, such as admissions requirements, student behavior, and financial aid, is presented in Table 9 through Table 11, respectively.

The target variable was completion with two classes: completer (finished in 150% time to completion) was identified by the number 1 and non-completer (did not finish in 150% time to completion) was identified by the number 0. In the preparation of the dataset, the completion variable, was defined as a multi-categorical variable (categories presented in Table 10) to specify students that would drop out or are still enrolled. In this manner policies for financial aid and overall GPA could be modeled. Once financial aid and GPA variables were created, the target variable was converted to binary by defining completers (less or equal than six years) and non-completers.

Table 9. Rules to define admission requirements data

Category: Secondary school
Admission requirements

Factor	Rules
ACT score	Mean 28, Standard deviation (STD) 1.73
High school GPA	Mean 3.56, STD 0.4183
Class rank	Mean 79, STD 19.2

Table 10. Rules to define student behavior data

Category: Student behavior	
Factor	**Rules**
Self-identified as having been dishonest	No 81% Yes 19%
Self-perceived ethicalness	Range 1 (not at all) to 7 (excellent) Mean 5.6, STD 1.2
Student belongs to a Greek fraternity/sorority	No 78% Yes 22%
Overall GPA	- Average GPA is 3.52, STD 0.28 - Minimum GPA for graduation is 2.0 - Minimum GPA for transfer is 2.25 - Minimum GPA for financial aid is 1.67
Degree completion	4 years 33% 6 years 24.6% 8 years 11.4% Transfer 25% Drop out 5% More than 8 years 1%

Table 11. Rules to define financial aid data

Category: Financial aid	
Factor	**Rules**
Received financial aid	Yes, need-based (NB) 27% Yes, non-need based (NN) 22% No, 51%

Once the dataset was created, a classification model for completion (target variable) was developed by applying the NN technique. STATISTICA 12 software was used in the implementation of the NN. The software uses an automated search that runs several networks with different combinations of initial parameters (e.g., training algorithm, number of hidden layers, error measure, and activation functions), next it retrieves the combinations with the highest classification accuracies. The model verification was performed using 10-fold cross-validation. The assessment of the model was performed based on the results presented in the confusion matrices and the overall accuracy of the resulting networks. The initial parameters of the networks are presented in Table 12.

Table 12. NN initial parameters

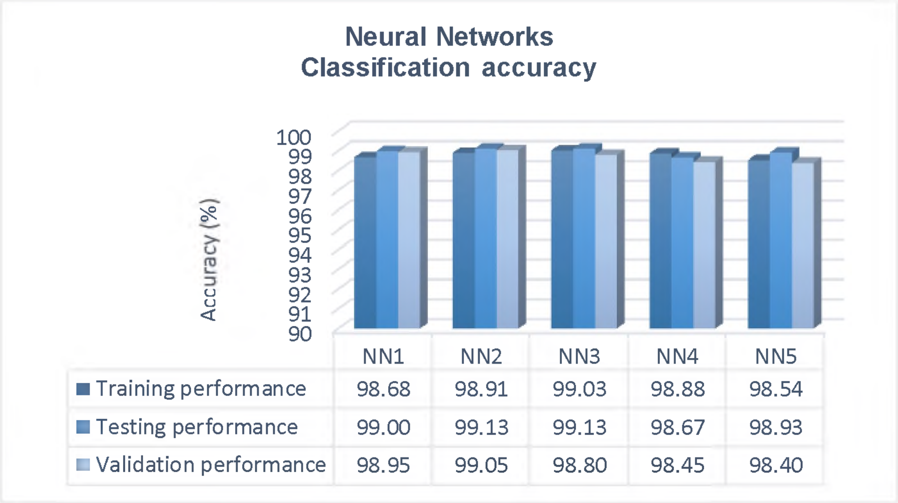
Parameter	MLP	RBF
Hidden units (min-max)	4-12	21-30
Activation and output functions	Exponential, hyperbolic tangent, logistic, identity, sin	
Error functions	Sum of Squares (SOS), entropy	
Number of networks generated	100	
Weight decay hidden and output	0.001-0.005	

3.3. MODEL RESULTS

Using the different combinations of the initial parameters, 100 networks were trained, tested, and verified. A summary of the best five performing models is presented in Table 13 and Figure 2.

Table 13. Results of best performing networks

Network ID	Training algorithm	Hidden layers	Training performance	Testing performance	Validation performance	Training cycles	Error function	Hidden activation function	Output activation function
1	MLP	6	98.6769	99.0000	98.9500	65	Entropy	Sine	Softmax
2	MLP	8	98.9077	99.1333	99.0500	23	Entropy	Tanh	Softmax
3	MLP	7	99.0308	99.1333	98.8000	44	SOS	Tanh	Tanh
4	MLP	9	98.8769	98.6667	98.4500	15	SOS	Exponential	Tanh
5	MLP	9	98.5385	98.9333	98.4000	19	SOS	Tanh	Sine



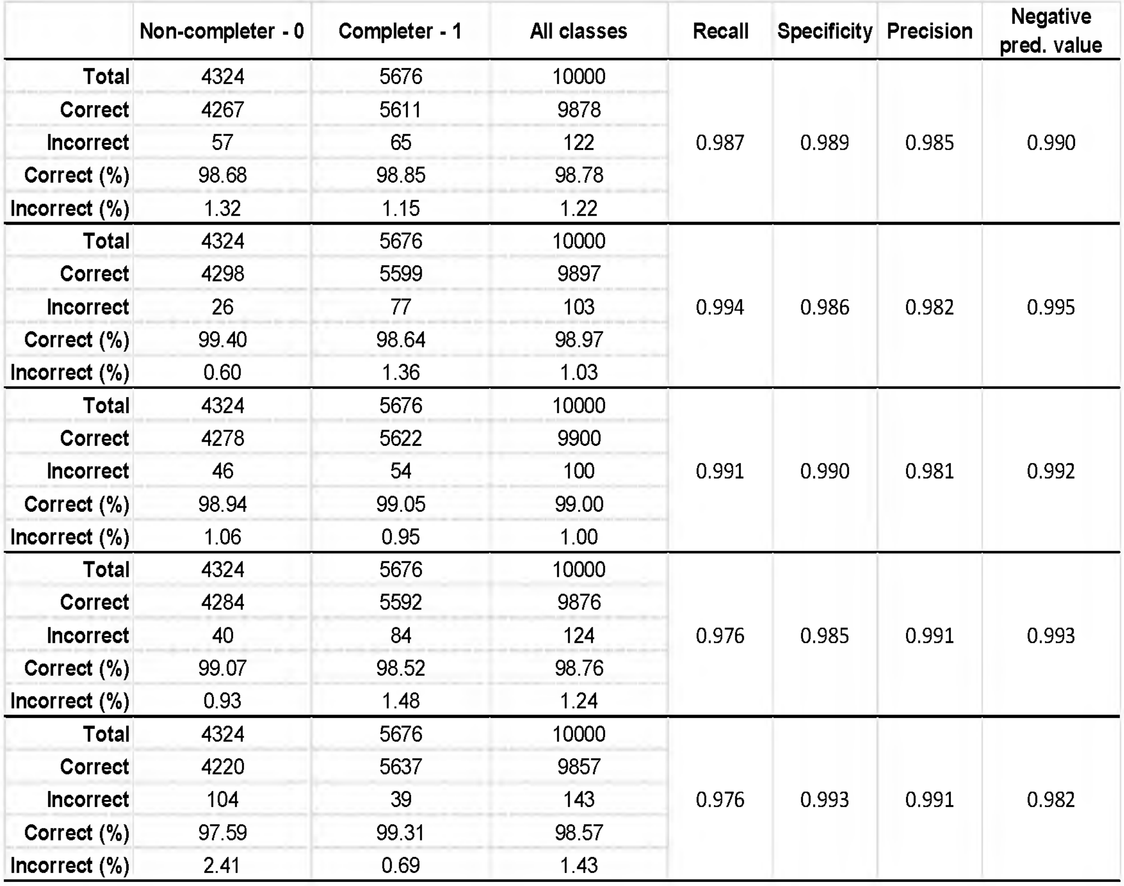
	NN1	NN2	NN3	NN4	NN5
■ Training performance	98.68	98.91	99.03	98.88	98.54
■ Testing performance	99.00	99.13	99.13	98.67	98.93
■ Validation performance	98.95	99.05	98.80	98.45	98.40

Figure 2. Training and validation classification performance

The results indicate high classification performance for the training, test, and validation sets. Further, the results indicate that every network is a good classifier of student success with relatively few misclassifications in each category (i.e., completer, and non-completer). Also, the classification accuracy for the validation sets in all networks is not significantly lower than for the training set, which is a positive sign that the networks were not overfitted.

The overall classification accuracy for the validation set is higher for model 2; however, when analyzing the classification summaries for each network and their assessment measures of recall, specificity, precision, and negative predictive value (Table 14), it was possible to conclude that network three (NN3) has the most consistent prediction behavior for both the completer and non-completer classes.

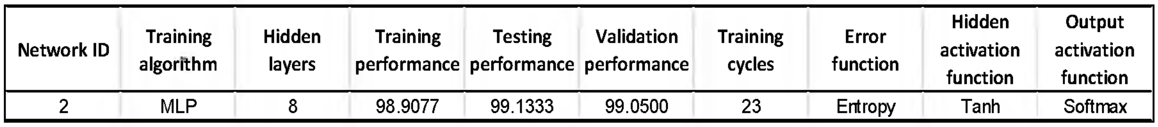
Table 14. Network classification summary



	Non-completer - 0	Completer - 1	All classes	Recall	Specificity	Precision	Negative pred. value
Total	4324	5676	10000	0.987	0.989	0.985	0.990
Correct	4267	5611	9878				
Incorrect	57	65	122				
Correct (%)	98.68	98.85	98.78				
Incorrect (%)	1.32	1.15	1.22				
Total	4324	5676	10000	0.994	0.986	0.982	0.995
Correct	4298	5599	9897				
Incorrect	26	77	103				
Correct (%)	99.40	98.64	98.97				
Incorrect (%)	0.60	1.36	1.03				
Total	4324	5676	10000	0.991	0.990	0.981	0.992
Correct	4278	5622	9900				
Incorrect	46	54	100				
Correct (%)	98.94	99.05	99.00				
Incorrect (%)	1.06	0.95	1.00				
Total	4324	5676	10000	0.976	0.985	0.991	0.993
Correct	4284	5592	9876				
Incorrect	40	84	124				
Correct (%)	99.07	98.52	98.76				
Incorrect (%)	0.93	1.48	1.24				
Total	4324	5676	10000	0.976	0.993	0.991	0.982
Correct	4220	5637	9857				
Incorrect	104	39	143				
Correct (%)	97.59	99.31	98.57				
Incorrect (%)	2.41	0.69	1.43				

Therefore, network three is the selected network for predicting student success using the specified variables to validate the architecture of the proposed system. Table 15 presents a summary of the network parameters.

Table 15. Network parameters of best performing network



Network ID	Training algorithm	Hidden layers	Training performance	Testing performance	Validation performance	Training cycles	Error function	Hidden activation function	Output activation function
2	MLP	8	98.9077	99.1333	99.0500	23	Entropy	Tanh	Softmax

The NN technique also allows for the identification of the impact of each variable in the model. It is calculated as a sensitive analysis of the error. STATISTICA 12 software tests the sensitivity of the error when simulating changes in the variables used in the network (e.g., if an important variable is removed the error will increase and vice versa). When the average error values from the different models is less than zero, the variable does not impact the model and can be removed.

As every NN has a different error, it is common to find slight changes in the order of impact of the variables for each model. Table 16 and Figure 3 presents the results for the rank of the variables for each model and the total rank is calculated as the average of all the results for each factor.

The most important predictors for this specific case in order of importance are ACT score, Class rank, and self-perceived ethicalness. The analysis also indicates that all the variables chosen for the model have some impact on the prediction.

Table 16. Variable rank from global sensitivity analysis

FACTOR **NN ID**	**NN 1**	**NN 2**	**NN 3**	**NN 4**	**NN 5**	**Average**
ACT score	1.849	4.197	2.663	2.050	3.715	2.895
Class rank	1.105	1.202	1.669	0.986	1.146	1.221
Self-perceived ethicalness	1.015	1.100	1.045	1.017	1.029	1.041
High school GPA	1.029	1.078	1.045	0.941	1.030	1.025
Financial aid	0.999	1.021	1.002	0.998	1.000	1.004
Greek student	1.030	0.964	1.004	0.991	1.027	1.003
Dishonesty	1.000	1.000	1.000	1.000	1.000	1.000
GPA	1.000	0.990	0.998	1.000	0.994	0.996

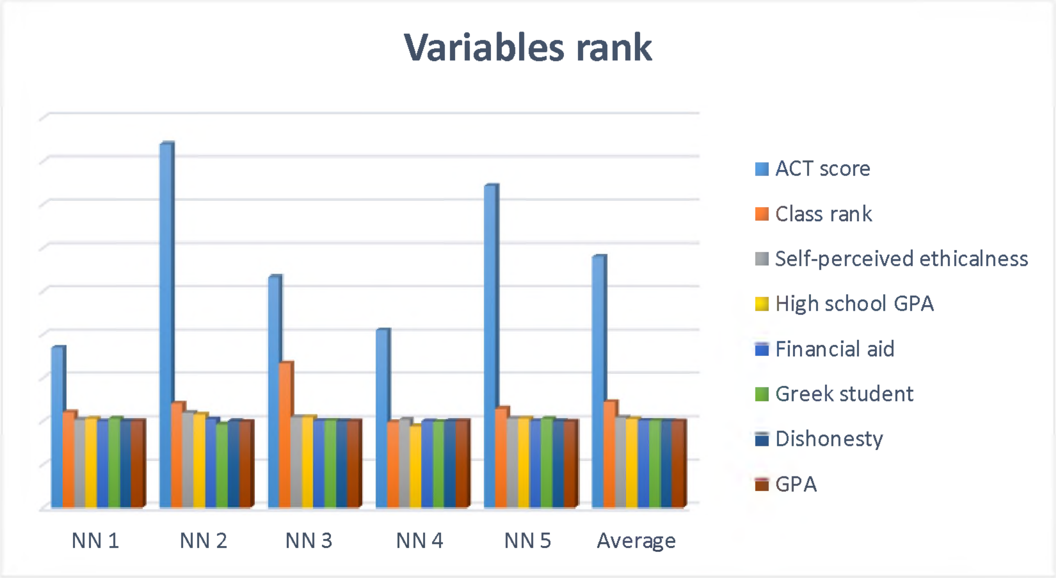


Figure 3. Variables rank

4. CONCLUSION, LIMITATIONS, AND FUTURE RESEARCH

The proposed system offers a clear picture of the interaction of the students' characteristics and their evolution through the course of the college experience. Further, it can be used as a base for formulating models that study student success.

This research proposed an architecture for a higher education system for student success, which was validated with promising results for the prediction of student completion. Using NN, the prediction accuracy obtained was above 98%. This work should be regarded as a preliminary effort to incorporate external and internal factors that impact the success of the students in higher education and how that emergent behavior can be predicted. Also, it continues demonstrating, as in prior literature, that the NN technique is an appropriate tool for student success prediction.

The results of the models elaborated from this system can enable the creation of strategies and reforms. The goal of those strategies and reforms would be increasing student completion rates. Lessons learned will also nurture the body of knowledge of accepted strategies and reforms that can be scaled and applied in other institutions.

One of the limitations to this study is the availability of statistical information concerning the different categories of factors included in the system architecture. Public information is limited, which reduced the scope of the validation to be only an effort to determine the effectiveness of the architecture without offering a good resource for the evaluation of improvement reforms. It is also important to note that the results of this study are not generalizable as they are specific to the institution studied. However, the proposed methodology can be applied to other institutions. Further, the analysis of the system could be evaluated by the design of a model that incorporates a more comprehensive set of factors. This study was limited to information and data that was previously collected and readily available.

It is key that institutions study student success models and strategies in a progressive manner, not only for small segments of the system. It should be based on holistic knowledge of how the system impacts students in their career journey. IT should be considered as the starting point to reconcile efforts to improve completion rates and success in general. To generate this support, there must an environment of trust, due to the sensitivity of the information and data that institutions can collect. Therefore, the reforms and strategies that can be formulated should be accompanied with the establishment of policies for evidence-based analytics that encompass the model-data transparency (collection and usage) to legal and ethical clarity.

REFERENCES

[1] Handel, S. J. (2013). The transfer moment: The pivotal partnership between community colleges and four-year institutions in securing the nation's college completion agenda. *New Directions for Higher Education*, *2013*(162), 5-15.

[2] Bailey, T. (2017). Community colleges and student success: Models for comprehensive reform. *Educause Review*, *52*(3), 33-42.

[3] Rosenbaum, J. E., Deil-Amen, R., & Person, A. E. (2007). *After admission: From college access to college success*. Russell Sage Foundation.

[4] Governor's Business Council, T. (2002). *Building An Effective and Aligned P-16 Education System: What Should Higher Education Do to Enhance Student Access and Success?*. ERIC Clearinghouse.

[5] Matthews, D. (2012). A Stronger Nation through Higher Education: How and Why Americans Must Achieve a Big Goal for College Attainment. A Special Report from Lumina Foundation. *Lumina Foundation for Education*.

[6] Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, *45*(1), 89-125.

[7] Tinto, V. (1999). Taking retention seriously: Rethinking the first year of college. *NACADA journal*, *19*(2), 5-9.

[8] Tinto, V. (2010). From theory to action: Exploring the institutional conditions for student retention. In *Higher education: Handbook of theory and research* (pp. 51-89). Springer, Dordrecht.

[9] Brock, T., Mayer, A. K., & Rutschow, E. Z. (2016). Using research and evaluation to support comprehensive reform. *New Directions for Community Colleges*, *2016*(176), 23-33.

[10] Restuccia, D., & Taska, B. (2018). Different skills, different gaps: Measuring and closing the skills gap. Developing Skills in a Changing World of Work: Concepts, Measurement and Data Applied in Regional and Local Labour Market Monitoring Across Europe, 207.

[11] Shapiro, D., Dundar, A., Huie, F., Wakhungu, P. K., Bhimdiwala, A., & Wilson, S. E. (2018). Completing college: A national view of student completion rates – Fall 2012 cohort (Signature Report 16). Herndon, VA: National Student Clearinghouse Research Center.

[12] Kirp, D. (2019). The college dropout scandal. The Chronicle Review. https://www.chronicle.com/interactives/20190726-dropout-scandal

[13] Jenkins, D., & Fink, J. (2015). What we know about transfer. New York, NY: Columbia University, Teachers College, Community College Research Center.

[14] Melguizo, T., Kienzl, G. S., & Alfonso, M. (2011). Comparing the educational attainment of community college transfer students and four-year college rising juniors using propensity score matching methods. *The Journal of Higher Education*, *82*(3), 265-291.

[15] Williford, A. M., & Schaller, J. Y. (2005, May). All retention all the time: How institutional research can synthesize information and influence retention practices. In Proceedings of the 45th Annual Forum of the Association for Institutional Research.

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

[17] Kelly, A. P., & Schneider, M. (Eds.). (2012). *Getting to graduation: The completion agenda in higher education*. JHU Press.

[18] Alkhasawneh, R., & Hargraves, R. H. (2014). Developing a hybrid model to predict student first year retention in STEM disciplines using machine learning techniques. Journal of STEM Education: Innovations and Research, 15(3), 35-42.

[19] Iam-On, N., & Boongoen, T. (2017). Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings. International Journal of Machine Learning and Cybernetics, 8(2), 497-510.

[20] Thomas, L., Herbert, J., & Teras, M. (2014). A sense of belonging to enhance participation, success and retention in online programs.

[21] Kondo, N., Okubo, M., & Hatanaka, T. (2017, July). Early detection of at-risk students using machine learning based on LMS log data. In Advanced Applied Informatics (IIAI-AAI), 2017 6th IIAI International Congress on (pp. 198-201). IEEE.

[22] Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. Expert Systems, 33(1), 107-124.

[23] Bahar Baran, & Eylem Kiḥç. (2015). Applying The CHAID Algorithm to Analyze How Achievement is Influenced by University Students' Demographics, Study Habits, and Technology Familiarity. Journal of Educational Technology & Society, 18(2), 323-335. Retrieved February 13, 2020, from www.jstor.org/stable/jeductechsoci.18.2.323

[24] Bailey, T., Jenkins, D., & Leinbach, T. (2005). Is Student Success Labeled Institutional Failure? Student Goals and Graduation Rates in the Accountability Debate at Community Colleges. CCRC Working Paper No. 1. *Community College Research Center, Columbia University*.

[25] Kruger, K., martin, R., Mehaffy, G., & O'Brien, J. (2017). Student Success: Mission-Critical. *Educause Review*, *52*(3), 11-20.

[26] Hiles, H. (2017). Student success, Venture Capital, and a Diverse Workforce: An Interview with Heather Hiles. *Educause Review*, *52*(3), 22-30.

[27] Bailey, T. (2017). Community colleges and student success: Models for comprehensive reform. *Educause Review*, *52*(3), 33-42.

[28] Klempin, S., & Karp, M. M. (2018). Leadership for transformative change: Lessons from technology-mediated reform in broad-access colleges. *The Journal of Higher Education*, *89*(1), 81-105.

[29] Susan Grajek and the 2016–2017 EDUCAUSE IT Issues Panel, "Top 10 IT Issues, 2017: Foundations for Student Success," EDUCAUSE Review 52, no. 1 (January/February 2017).

[30] Dahlstrom, E. (2016). Moving the red queen forward: Maturing analytics capabilities in higher education. *Educause Review*, *51*(5), 36-54.

[31] Birta, L. G., & Arbez, G. (2013). *Modelling and simulation*. London: Springer.

[32] Wasson, C. (2019, July). The State of Systems Engineering Technical Practice versus Discipline: A Survey of INCOSE Chapter Attendees in North America. In *INCOSE International Symposium* (Vol. 29, No. 1, pp. 591-619).

[33] Dawood, M., Tapia, J., Trujillo, K., Guynn, M., & Wojahn, P. (2017, July). Preliminary results on students' study habits and their grades in STEM courses. In 2017 USNC-URSI Radio Science Meeting (Joint with AP-S Symposium) (pp. 25-26). IEEE.

[34] Hora, M. T., & Oleson, A. K. (2017). Examining study habits in undergraduate STEM courses from a situative perspective. International Journal of STEM Education, 4(1), 1.

[35] Alban, M., & Mauricio, D. (2019). Predicting university dropout through data mining: a systematic literature. *Indian Journal of Science and Technology*, *12*(4), 1-12.

[36] Kuh, G. D., Kinzie, J. L., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2006). *What matters to student success: A review of the literature* (Vol. 8). Washington, DC: National Postsecondary Education Cooperative.

[37] Cardona, T., Cudney, E., Snyder, J., & Hoerl, R. (2020). *Data mining and machine learning retention models in higher education, a systematic review* [Manuscript submitted for publication]. Department of Engineering Management and Systems Engineering, Missouri University of Science and Technology

[38] Cardona, T., Cudney, E., & Snyder, J. (2019a). Predicting degree completion through data mining. Proceedings of the ASEE Annual Conference & Exposition, Tampa, FL.

[39] Cardona, T., Cudney, E., & Snyder, J. (2019b). Predicting student retention using artificial neural networks. Proceedings of the IISE Annual Conference, Orlando, FL.

[40] Cardona, T., & Cudney, E. (2019). Predicting student retention using support vector machines. Proceedings of the International Conference on Production Research Manufacturing Innovation: Cyber Physical Manufacturing, Chicago, IL.

[41] Cardona, T., Cudney, E., Snyder, J., & Hoerl, R. (in press). Predicting student degree completion using random forest. Proceedings of the ASEE Annual Conference & Exposition.

[42] Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. Decision Support Systems, 49(4), 498-506.

[43] Delen, D. (2011). Predicting student attrition with data mining methods. Journal of College Student Retention: Research, Theory & Practice, 13(1), 17-35.

[44] Dissanayake, H., Robinson, D., & Al-Azzam, O. (2016, January). Predictive Modeling for Student Retention at St. Cloud State University. In Proceedings of the International Conference on Data Mining (DMIN) (p. 215). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[45] Babić, I. D. (2017). Machine learning methods in predicting the student academic motivation. Croatian Operational Research Review. 8, 443-461.

[46] Raju, D., & Schumacker, R. (2015). Exploring student characteristics of retention that lead to graduation in higher education using data mining models. Journal of College Student Retention: Research, Theory & Practice, 16(4), 563-591.

[47] Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. Journal of Applied Research in Higher Education, 10(1), 61-75.

III. PREDICTING DEGREE COMPLETION THROUGH DATA MINING

Tatiana A. Cardona[a], Elizabeth A. Cudney[a], and Jennifer Snyder[b]

[a]Deparment of Engineering Management and Systems Engineering, Missouri University of Science and Technology

b School of Science, Valencia College's East campus.

ABSTRACT

Universities and colleges continuously strive to increase student retention and degree completion. The U.S. Department of Education has set the goal of preparing a society with individuals capable to "understand, explore and engage with the world" specific skills that can be achieved through STEM majors. Currently, considerable student data are collected and there is a latent opportunity to make the available information useful for determining the factors that influence retention and completion rates. Analyzing student data with those aims is vital for intentional student advising. To this end, this research presents the application of decision trees to predict degree completion within three years for STEM community college students. Decision trees also enable the identification of the factors that impact program completion using non-parametric models by classifying data using decision rules from the patterns learned. The model was developed using data on 283 students with 14 variables. The variables included age, gender, degree, and college GPA, among others. The results offer important insight into how to develop a more efficient and responsive system to support students.

Keywords: Student retention, decision trees, degree completion, engineering education

1. INTRODUCTION

One of the main concerns for universities and colleges is attrition rate. Students able to complete their degrees in the expected time directly impacts the reputation of the institution, as it reflects institutional commitment on contributing to the society by preparing individuals capable of engaging with the world (Williford & Schaller, 2005). Despite this, retention rates are currently low. With respect to college and university students pursuing STEM majors, retention rates are 69% and 48%, respectively (Snyder & Cudney, 2018). Colleges and universities collect considerable student data. However, their ability to process the available information does not occur at the same pace as the collection (Morris, 2016). Therefore, effort needs to be made on making the data useful to improve student retention. For instance, by determining the factors that influence student retention and completion rates, it is possible to improve the intentional student advising, planning, and development of retention strategies based on student needs (Slim et al., 2005). In recent years machine learning techniques have been applied to process educational data, which aligns with the focus on improving the processing of information. According to the literature, those techniques offer predictions of student dropout with high confidence (Pereira & Zambrano, 2017). Within machine learning techniques, decision trees (DT) have been employed successfully to predict and classify factors that impact student success measured as risk of dropout, attrition risk, and completion risk. The purpose of this research was to develop a prediction model to forecast program completion within three years by STEM community college students and identify the factors that influence successful completion. To this end, this paper presents the application of DT as a machine learning technique using a data base comprised of 283 entries with 14 variables collected from a community college in the Midwest. DT was used to develop a predictive model for student success. The key research question is: Can DT accurately predict student completion rates? The remainder of this paper is structured into the following sections: literature review and background on DT applications on student success prediction, research methodology, results, and conclusions and future work.

2. LITERATURE REVIEW

DT have been one of the most frequently applied machine learning techniques for prediction of student success and identification of factors that influence it. According to Adejo and Connolly (2018), the advantage of DT resides on the computational speed and flexibility for modelling nonlinearity. Further, DT structures are easy to understand and communicate; however, the main weakness is the overfitting/underfitting with an option to mild it by pruning. Several studies reflect the idea that DT offered a more visual structure of the results and state the importance of using the technique although other techniques could have better accuracy results (Delen, 2010; Delen, 2011; Oztekin, 2016). Research by Delen (2010, 2011) found that the classification of factors indicated that fall GPA, loans, and financial aid had a significant impact on predicting student attrition. Oztekin (2016) developed a hybrid method to predict completion for undergraduate students and also found that GPA was an important predictor variable. Several studies applied principal component analysis (PCA) to a data set to filter the number of variables to be included in the model (Dissanayake et al., 2016; Adejo and Connolly, 2018). In the study by Dissanayake et al. (2016), not all techniques showed improvement in the results when applying PCA. Rather, DT showed better performance when using the original dataset. In another study, Babić (2017) developed a classification model for predicting student academic motivation. The methodology included the application of machine learning classifiers such as neural network (NN), DT, and support vector machine (SVM). The results showed there was not a significant difference in the performance of the techniques. Supporting this conclusion Miranda and Guzman (2017) identified the factors that determine student dropout by applying different data mining techniques including Bayesian network classifier, DT, and NN. The results showed there was no significant difference within the performance of each technique. Additional comparison of methods to identify key factors that impact the accuracy of an earlyalert system was conducted to determine the level of factor importance. Pereira and Zambrano (2017) identified that the most relevant academic factors were low average in grades, number of failed classes in initial semesters, and department of study. Further, the relevant socioeconomic factors were university enrollment fee and provenance from south of the department. While, Tsao et al. (2017) concluded that the variables chosen for creating the datasets greatly impact the performance of the prediction models. Uddin and Lee (2017) developed a hybrid model to predict a good fit in major for students to decrease dropout risk. Two algorithms that used several machine learning techniques including DT were integrated in the master algorithm to quantify the academic success factor. The results evidenced that the more data the more accurate the prediction. The hybrid method outperformed several known stand-alone techniques. The DT methodology has been successfully used to predict academic success in higher education. However, most of the research has been performed in universities, rather than community colleges. The lack of research is this area indicates that more research should be performed to increase retention and completion of STEM students in community colleges

3. RESEARCH METHODOLOGY

The data utilized for this research was collected from a community college located in Missouri. The community college offers associates degrees in STEM fields. Further, the community college allows students to declare their major upon entrance, which makes it ideal for data analysis. The data was collected for five years. The research process was conducted in the following stages: 1) data description and preparation, 2) data modeling and application of DT, and 3) model assessment. A pictorial representation of the modeling process is provided in Figure 1. The stages are explained in more detail in the following subsections.

3.1. DATA PREPARATION

The data for this research was collected from a community college in the Midwest, which offers associate degrees in STEM majors. The dataset was comprised of five years of registered students, which consists of 904 students pursuing degrees in chemistry, biology, and engineering. From this data, 177 were identified as completing the degree within three years (150% of normal time for completion as required to be reported by the 1990 Student Right-to-Know Act for postsecondary institutions). The remaining 727 students did not graduate within that period, which is most commonly due to college withdrawal or switching to a non-STEM major. The data set was cleaned because of considerable missing and inconsistent data. For example, standardized exam scores were not available or provided for some students. After cleaning the data from incomplete records, a final dataset of 282 students was selected, which consisted of 51 completers and 231 non-completers. The data set had 14 variables, a non-exhaustive number for computational purposes. These variables were selected as they were readily collected and available. Therefore, it was not necessary to reduce the number of variables on the data. Table 1 provides a list of the variables used in the research.

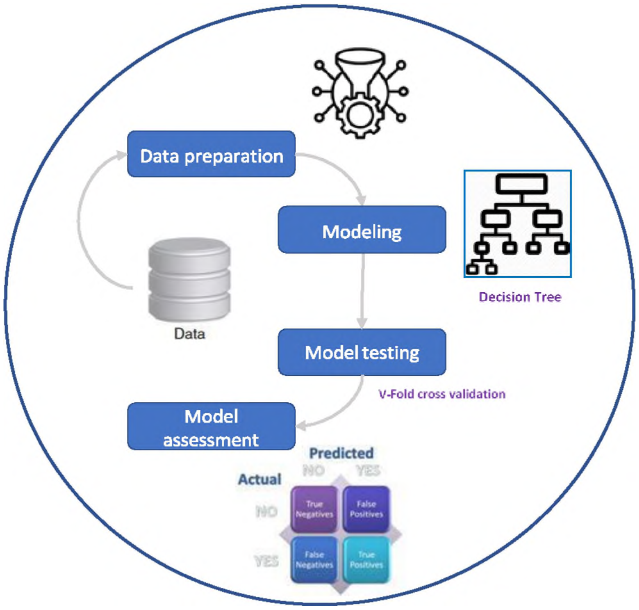


Figure 1. Data analytic methodology

Table 1. Variables used in the study

Variable	Type
Complete (Target variable)	Yes/No
Degree	Chemistry, Biology, Engineering
Age	Numerical
Gender	Female/Male
Full Time Student	Yes/No
1st Generation Student	Yes/No
Plan to work	Yes/No
ACT comprehensive	Numerical
ACT English	Numerical
ACT mathematics	Numerical
ACT reading	Numerical
High school GPA	Numerical
College GPA	Numerical

3.2. DATA MODELING

A DT is a tree like structure with a hierarchical nature. It can visually represent a decision-making process that divides the data as univariate splits for categorical predictor variables. The goal of DT is the prediction on a dependent variable, but also variable classification can be done by using this technique. The structure consists of classes (leaves), attributes (internal nodes), and connecting attributes (branches). It traces the path of nodes and branches to generate the prediction. DT are flexible in the fact that they examine the effects of the predictor variable one at time and can be computed for categorical and numerical predictors (Breiman et al., 1984).

In this study, classification, and regression tree technique (CART) was used. This method for splitting selection generates an exhaustive search for univariate split producing the maximum goodness of fit. The stopping criteria selected was FACT. It allows for splitting until nodes contain no more cases than a specified fraction of the size of the class. For this study, 0.05 was the fraction used. It was also important to set the model to be equally precise for predicting students that could complete on time as for predicting the ones who could not. A cross validation of 10 folds was set in the training and a global cross validation was generated after running the training to validate the model. The model was implemented using Statsoft Statistica 12.

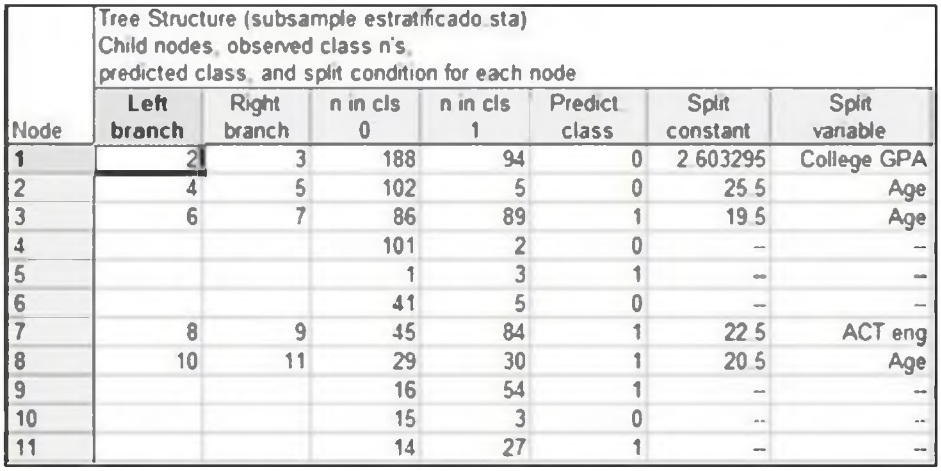
3.3. MODEL ASSESSMENT

The model was assessed using measures of performance in training and the misclassification matrix. For testing the prediction, a 10-fold global cross validation was generated, and the results were compared with the cross validation generated with the training. The overall performance is calculated as the proportion of correctly classified values from the sample size (N). For the identification of factors that impact the prediction, Statsoft Statistica 12 presents the results for predictor importance as a table with a ranking score in a range of 0-100 for each predictor.

4. RESULTS

The selected tree had 11 nodes, within 6 are terminal nodes. The results are presented in Table 2 and Figure 2. Prediction class is 1 for completer or 0 for non-completer. Terminal nodes 4, 6, and 10 had a prediction of non-completer with 2, 5, and 3 misclassifications, respectively. While terminal nodes 5, 9, and 11 had prediction of completer with 1, 16, and 14 misclassifications, respectively. College GPA, age, and ACT Engineering were used as the splitting variable.

Table 2. Selected tree results



Tree Structure (subsample estratificado.sta)
Child nodes, observed class n's,
predicted class, and split condition for each node

Node	Left branch	Right branch	n in cls 0	n in cls 1	Predict class	Split constant	Split variable
1	2	3	188	94	0	2.603295	College GPA
2	4	5	102	5	0	25.5	Age
3	6	7	86	89	1	19.5	Age
4			101	2	0	--	--
5			1	3	1	--	--
6			41	5	0	--	--
7	8	9	45	84	1	22.5	ACT eng
8	10	11	29	30	1	20.5	Age
9			16	54	1	--	--
10			15	3	0	--	--
11			14	27	1	--	--

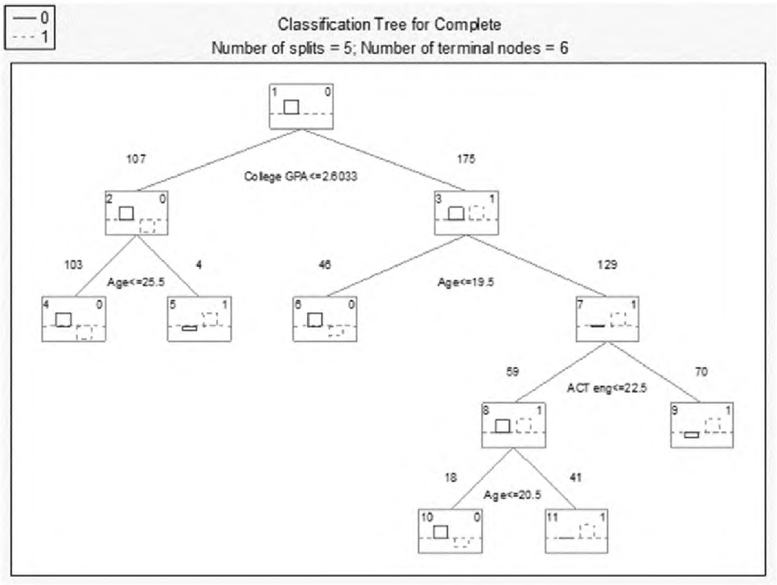
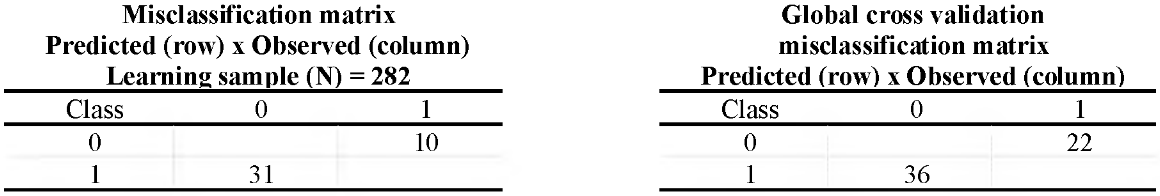


Figure 2. Selected tree

The cost matrices from the training and test data are displayed in Table 3. The overall performance for the training and testing is consistent with not a significant difference (85.47% and 79.43%, respectively). The cross validation was also evaluated to ensure the consistency. Therefore, training cross validation cost and global cross validation cost and their respective standard deviations were compared for similarities (Table 4). In conclusion, the cost percentages in training and testing are very similar, which confirms consistency on the predictions.

Table 3. Misclassification matrix. Left, training data. Right, testing data



Misclassification matrix
Predicted (row) x Observed (column)
Learning sample (N) = 282

Class	0	1
0		10
1	31	

Global cross validation
misclassification matrix
Predicted (row) x Observed (column)

Class	0	1
0		22
1	36	

Table 4. Results statistics. Left, training. Right, testing

Training tree statistics

CV cost	0.1985
Std	0.0251

Test tree statistics

CV cost	0.2057
Std	0.0241

The results indicate that the DT methodology offers a good prediction model for STEM degree completion for community college students with the specified variables with validation performance of approximately 80%.

After evaluating the prediction abilities of the model, it was important to identify the variables that impact the prediction. Table 5 presents the classification of level of importance of the different predictors. The results showed Figure 3 that the most significant variables are college GPA, age, ACT math, and ACT English.

Table 5. Predictor importance

Variable	Ranking
Gender	2
Full time student	19
Part time student	8
First generation	2
Plans to work	15
Degree	13
ACT Comprehensive	43
ACT English	48
ACT Mathematics	53
ACT Reading	31
High School GPA	43
College GPA	59
Age	100

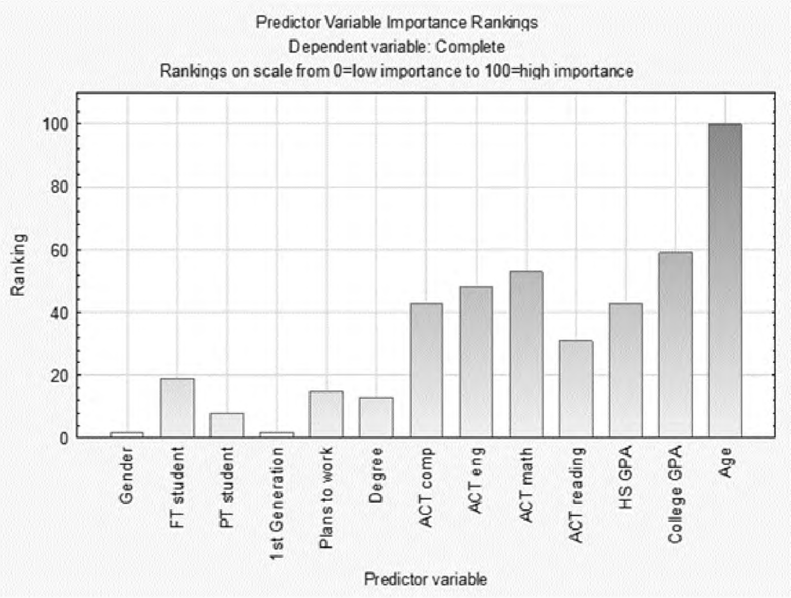


Figure 3. Predictor importance

5. CONCLUSIONS AND FUTURE WORK

This research presented a complete case of applying DT, which indicates that it is an effective tool for forecasting completion success of community college students in

STEM majors. Also, it can be used for identifying the level of importance of the factors impacting such prediction. Although GPA is a common factor founded in prior literature as important for the prediction of student success, variables such as ACT math and ACT English are not commonly found in other studies. This statement infers what was found in the literature in terms of the variables chosen for the model impact its performance. Also, the findings suggest that the level of importance of those factors depended on the methodology used; however, further investigation should be performed.

As with any research study, there are limitations. First, the research findings are not generalizable as the study was conducted on data from only one community college. In addition, community colleges are representative of their local demographics. Therefore, results from one community college will not be generalizable to another university. However, the methodology should be applicable for the analysis. Next, the research was conducted using available data. The community college had information only on 14 variables. Numerous additional variables were identified through the literature. Future research should utilize data collected using considerably more data as noted in the relevant literature.

Further studies can also focus on combining a more complete mixture of factors to have a more robust model. In that manner a prediction model with the right set of variables can represent a useful tool for the creation of retention strategies by addressing the advising.

REFERENCES

Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education, 10*(1), 61-75.

Babić, I. D. (2017). Machine learning methods in predicting the student academic motivation.

Croatian Operational Research Review. *8*, 443-461.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees, Chapman and Hall/CRC Press, Boca Raton, Florida.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498-506.

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice*, *13*(1), 17-35.

Dissanayake, H., Robinson, D., & Al-Azzam, O. (2016, January). Predictive Modeling for Student Retention at St. Cloud State University. In *Proceedings of the International Conference on Data Mining (DMIN)* (p. 215). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Miranda, M. A., & Guzmán, J. (2017). Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos. *Formación universitaria*, *10*(3), 61-68.

Morris, L. V. (2016). Mining Data for Student Success. *Innovative Higher Education*, *41*(3), 183- 185.

Oztekin, A. (2016). A hybrid data analytic approach to predict college graduation status and its determinative factors. *Industrial Management & Data Systems*, *116*(8), 1678-1699.

Pereira, R. T., & Zambrano, J. C. (2017, December). Application of Decision Trees for Detection of Student Dropout Profiles. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on* (pp. 528-531). IEEE.

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

Snyder, J., & Cudney, E. A. (2018, June), *A Retention Model for Community College STEM Students* Paper presented at 2018 ASEE Annual Conference & Exposition, Salt Lake City, Utah. https://peer.asee.org/29719

Tsao, N. L., Kuo, C. H., Guo, T. L., & Sun, T. J. (2017, July). Data Consideration for At-Risk Students Early Alert. In *Advanced Applied Informatics (IIAI-AAI), 2017 6th IIAI International Congress on* (pp. 208-211). IEEE.

Uddin, M. F., & Lee, J. (2017). Proposing stochastic probability-based math model and algorithms utilizing social networking and academic data for good fit students prediction. *Social Network Analysis and Mining*, *7*(1), 29.

Williford, A. M., & Schaller, J. Y. (2005, May). All retention all the time: How institutional research can synthesize information and influence retention practices. In *Proceedings of the 45th Annual Forum of the Association for Institutional Research*

IV. PREDICTING STUDENT RETENTION USING ARTIFICIAL NEURAL NETWORKS

Tatiana A. Cardona[a], Elizabeth A. Cudney[a], and Jennifer Snyder[b]

[a]Deparment of Engineering Management and Systems Engineering, Missouri University of Science and Technology

b School of Science, Valencia College's East campus.

ABSTRACT

Universities and colleges continuously strive to increase student retention and degree completion as they are directly related with university rankings by measuring institutional performance and success. In addition, the U.S. Department of Education has set the goal of preparing a society with individuals capable to "understand, explore and engage with the world", which are specific skills that can be achieved through STEM majors. To achieve these objectives, colleges and universities collect vigorous amounts of student data. Analyzing student data is vital to determining the factors that influence student retention and completion rates by providing insight into opportunities for intentional student advising. To this end, this research presents the application of artificial neural networks (ANN) to predict degree completion within three years by STEM community college students. ANN enables the classification of the input variables into expected results, retention, and completion, by learning from the error produced by the model and adjusting the weights of the input variables. The model was developed using data on 283 students with 14 variables. The variables included age, gender, degree, and college GPA, among others. The model results, which include prediction and variables ranking, offer an important understanding about how to develop a more efficient and responsive system to support students.

Keywords: Student retention, neural networks, degree completion, engineering education.

1. INTRODUCTION

Colleges and universities collect vigorous amounts of data from students from as soon as they apply to the institution. The improvement in processing that data is vital to obtain positive gains about the factors that influence degree completion rates. To this end, the prediction or forecasting of program completion by the student gives insight into areas in need of development to improve advising according to Zhang et al. (2004) and increase retention and graduation rates. As defined by the National Center for Education Statistics (NCES) in 2018, retention rates refer to the proportion of students returning to the same institution the following fall, while graduation rates are students who complete the programs in certain amount of time according to. These terms were adopted for the development of the present study.

The U.S. Department of Education set a goal of preparing a society with individuals capable to "understand, explore and engage with the world", which are specific skills that can be achieved through STEM majors. However, as presented in Morris (2016) retention rates for college and university students pursuing STEM majors are low, 69% and 48%, respectively. Thus, the literature indicates an important interest on increasing student retention in higher education as it reflects institutional commitment to the students (Snyder and Cudney, 2018). Therefore, determining the factors that influence student retention and completion rates provides insight into opportunities for intentional student advising, better planning, and development of retention strategies based on student needs (Williford and Schaller, 2005).

Recently, machine learning techniques have been applied to process educational data focused on student success measured as risk of dropout, attrition risk, and completion risk, which translates to retention and graduation rates (Williford and Schaller, 2005). Neural networks (NN) have been employed to predict and classify factors that impact such measures. Within the models in the current literature, NN have proven to have superior performance than other machine learning techniques based on prediction accuracy.

The purpose of this research was to develop a prediction model to forecast program completion within three years by STEM community college students. Further, the factors that influence successful completion were identified and compared to prior research using the same data with different methods specifically Snyder and Cudney (2018) Therefore, the current focus is on information processing or, in other words, on the need of generating models that help to make the available data useful (Slim et al., 2014). To this end, this paper presents the application of NN as a machine learning technique using a database comprised of 283 entries with 14 variables collected from a community college in the Midwest.

The remainder of this paper is structured into the following sections: literature review and background on NN applications on student success prediction, data analysis and predictive model development and validation, and comparison to prior research.

2. LITERATURE REVIEW

Several machine learning techniques have been applied to generate prediction models and identify the factors that influence retention and graduation rates in higher education. One of the most widely used techniques is NN. The structure of NN consists of an input layer of neurons, one or more hidden layers, and an output layer. As explained in Hassoun (1995) and in Haykin (2009), layers are connected in a forward manner, i.e. adjacent layers are fully interconnected by weights in the first layer, and activation functions in the following layers to generate the outputs. The learning process consists of changing the weights on the training dataset to decrease the prediction error.

NN has been effectively applied to forecast student success as several studies showed performance in prediction over 70% classifying it as one of the most effective methods. For instance, Alkhasawneh and Hargraves (2014) developed a hybrid model to predict first year retention in STEM majors. The research was divided into a qualitative and a quantitate stages to further construct a hybrid model. NN was used for modeling and an accuracy of 79% was obtained in the predictions.

Babić (2017) made a comparison of techniques; however, the results from comparing techniques (NN within them) were not different through applying a test of significance. Therefore, their efficiency was evaluated based on their capacity to predict academic motivation using analysis of the confusion matrix. From this evaluation, NN had a better prediction performance. The research found that NN with a radial basis function (RBF) was the most efficient method to predict below-average academic motivation with a 100% negative predictive value. In a similar study, Miranda and

Guzman (2017) found there was not significant difference between the prediction models used.

Data preparation is an important step for the application of machine learning techniques mostly when using unbalanced datasets. For example, Delen (2010), and Delen (2011) compared four different machine learning techniques to predict student success. The findings indicated that machine learning techniques, specifically NN and support vector machines (SVM), have a better performance when working with a balanced dataset. Other studies also undergo a cleaning and balanced process before applying NN and other machine learning techniques (Oztekin, 2016 and Adejo, 2018).

In terms of variables selection for studies focused on student success, high school GPA and ACT composite scores are important factors to include in prediction models according to Radunzel and Noble (2012) and Schmitt et al. (2009). For studies that used specifically NN as the prediction methodology, this statement continues to be true as several studies identified academic factors (including freshman GPA, high school GPA, ACT and SAT scores) and financial situation as good predictors as found in Miranda and Guzman (2017), Delen (2010) and Delen (2011). Further, for institutions with primarily STEM majors, ACT math, prior science preparation, and gender influenced student success (Alkhasawneh and Hargraves, 2014). However, the data used in each study has a different combination of factors that can represent different levels of ranking.

3. RESEARCH METHODOLOGY

The research process was conducted in the following stages: 1) data description and preparation, 2) data modeling, application of NN, and 3) model assessment and comparison of results with prior study. A pictorial representation of the modeling process is depicted in Figure 1. The stages are explained in more detail in the following subsections.

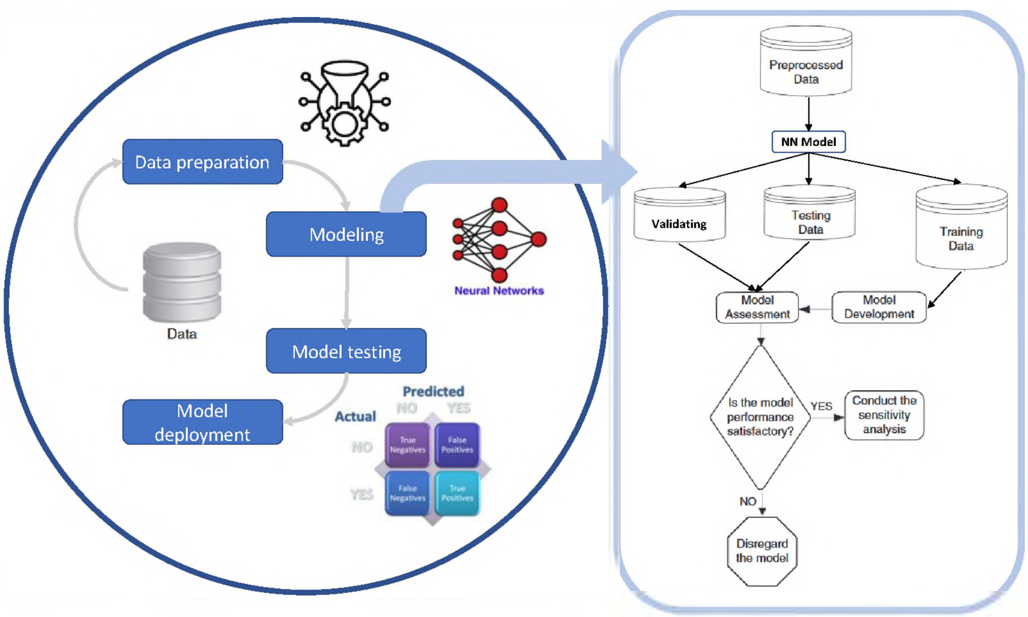


Figure 1. Data analytic methodology

3.1. DATA DESCRIPTION AND PREPARATION

The data for this research was collected from a community college in the Midwest, which offers associate degrees in STEM majors. The database was previously processed in a separate research study (Snyder and Cudney, 2018). The treatment of the data in the first stage was performed using the same process as the prior study to ensure consistency when comparing results from the different methodologies.

The dataset was comprised of five years of registers from 904 students pursuing degrees in chemistry, biology, or engineering. From this data, 177 were identified as completing the degree within three years, which is 150% of normal time for completion as required to be reported by the 1990 Student Right-to-Know Act for postsecondary institutions. The remaining 727 students did not graduate within that period with reasons considered as of college withdrawal or switching to a non-STEM major. According to Snyder and Cudney (2018), the data set had to be cleaned because of considerable missing data and inconsistent data; for example, standardized exam scores were missing for some students as this information is not required for community college admission. After cleaning the data to remove incomplete records, the final dataset consisted of 282 students, which consisted of 131 non-completers and 51 completers.

For the present study, reducing the number of variables on the data was not necessary before running the NN model. The number of variables resulting after cleaning the data was moderate for developing the network, which later would be able to classify the variables by level of importance in the prediction model. Table 1 provides a list of the variables used in the research.

3.2. DATA MODELING

NN are powerful analytical techniques inspired by the functionality of the brain. Although NN provides a loose approximation, it uses a process structured based on animal neurons and can predict new observations from old observations using an iterative learning process. It enables the classification of the input variables into expected results (output) by learning from the error produced by the model and adjusting the weights of the input variables to improve the predictions. The network trains to reduce the error. NN can be applied to categorical and numerical data. A key advantage of NN is it is suitable to work with nonparametric models making it more flexible to replicate reality (Haykin, 2009).

Table 1. Variables used in the study

Variable	Type
Complete (Target variable)	Yes/No
Degree	Chemistry, Biology, Engineering
Age	Numerical
Gender	Female/Male
Full Time Student	Yes/No
1st Generation Student	Yes/No
Plan to work	Yes/No
ACT comprehensive	Numerical
ACT English	Numerical
ACT mathematics	Numerical
ACT reading	Numerical
High school GPA	Numerical
College GPA	Numerical

In this study, multilayer perceptron (MLP) and (RBF) networks were used. Both methods consist of inputs, hidden layers, and output layers. The difference is found in the input-target relationship. MLP network models relate input data to the target in one stage using the weights. While RBF network performs this in two stages: 1. models' probability of input data using the RBF (location and radial spread) and 2. Relates the input data to the target (weights).

The model was implemented using Statsoft Statistica 12. The parameters for training the models were set as shown in Table 2. The modeling was set on automated network search (ANS) mode in STATISTICA software. This option allows optimum models to be determined within the cycles programmed.

Table 2. Modeling parameters

	Parameter/ Model	**MLP**	**RBF**
Variable	Activation functions	Exponential, hyperbolic tangent, logistic	
	Error functions	Sum of squares (SOS), cross entropy (CE)	
	Hidden units (min-max)	5-25	10-30
Fixed	Training cycles	200	

3.3. MODEL ASSESSMENT AND COMPARISON

The model was assessed using measures of performance in training, test, and validation. Also, recall and recall measures were analyzed based on the confusion matrix. The last step ensures a more holistic analysis of the results by mitigating possible misinterpretations. It is important for the model to be as precise for predicting students that could complete on time as for predicting the ones who could not.

Determining the level of importance of variables used in the model was done with a global sensitivity analysis. The results were compared with the factors found important for the model in a previous study that used the same data set and Mahalanobis Taguchi System and regression models (Snyder and Cudney, 2018). The comparison allows for conclusions on the behavior of the data through different algorithms and performance of the models.

Initial experiments showed high performance in prediction model but low recall, which could be attributed to the unbalance in the data for the target value (more completers than non-completers). Therefore, it was necessary generate a subsample to balance the number of instances for both classes. This was done using the stratified sampling function in STATISTICA 12. To ensure consistency in the new sample, an ANOVA analysis was conducted to confirm there was no change in the means of the numerical variables. The results showed there was no significant difference between the means with a *p*-value of 0.9638.

4. RESULTS

The NN application generated 10 models that were selected for evaluation of recall and overall performance (accuracy). All models are MLP type and showed to be efficient with overall performance measures over 85% for training data, over 88% for testing data, and over 83% for validation data. The results are presented in Table 3.

Table 3. Networks with better performance

	Performance (%)		
Network	**Training**	**Test**	**Validation**
MLP 1	85.86	90.48	83.33
MLP 2	95.96	92.86	88.10
MLP 3	94.95	95.24	85.71
MLP 4	96.46	92.86	85.71
MLP 5	88.38	90.48	85.71
MLP 6	92.42	92.86	88.10
MLP 7	90.40	90.48	85.71
MLP 8	95.96	95.24	88.10
MLP 9	92.40	95.24	85.71
MLP 10	93.94	88.1	83.33

Based on the validation, the models with highest performance are MLP2, MLP6, and MLP 8 as illustrated in the confusion matrix in Table 4. When proceeding to the evaluation of recall for the three selected models, MLP 8 was determined to have a better prediction for both classes (completer 96.32% and non-completer 95.16%). Although, it is important to consider that the unselected models presented a high ability for prediction with recall measures over 93%. The selected model offers a more balanced recall output. From the results is evident that NN methodology offers a good prediction model for STEM degree completion for community college students with the specified variables.

After evaluating the prediction abilities of the model, it was important to identify the variables with more impact in the prediction. With this aim, a sensitivity analysis was conducted. STATISTICA tests the sensitivity of the error when simulating changes in the variables used in the network, e.g. if an important variable is removed the error will increase and vice versa. When the average error values from the different models is less than zero the variable does not impact the model and can be removed. The results showed that the most significant variables are full time student, first generation, degree, and college GPA as shown in Table 5. The analysis also indicates that all the variables chosen for the modeling have some impact in the prediction.

4.1. COMPARISON OF RESULTS WITH PREVIOUS STUDY

The identification of factors that influence student success on completing STEM degrees is equivalent to the prior study performed with the same dataset where the most significant factors were college GPA, full time student, and gender (Snyder and Cudney, 2018). This consistency supports the idea that the modeling technique does not impact in a significant manner the level of the importance of factors influencing completion of STEM majors specifically for community college students, nevertheless, further research be conducted by applying other techniques to confirm this statement.

Table 4. Confusion matrix for the selected models

MLP 2

Class	Completer	Non-completer	Total
Completer	132	4	136
Non-completer	4	58	62
Total	136	62	198
Correct (%)	97.06%	93.55%	
Incorrect (%)	2.94%	6.45%	

MLP 6

Class	Completer	Non-completer	Total
Completer	132	4	136
Non-completer	4	58	62
Total	136	62	198
Correct (%)	97.06%	93.55%	
Incorrect (%)	2.94%	6.45%	

MLP 8

Class	Completer	Non-completer	Total
Completer	131	5	136
Non-completer	3	59	62
Total	134	64	198
Correct (%)	97.76%	92.19%	
Incorrect (%)	2.24%	7.81%	

In terms of prediction performance and accuracy (see Table 6), the model in the prior literature and the one in the present study can be consider effective as they can generate predictions with performance over 80%. Revising the percentages, NN showed a more balanced accuracy when correctly predicting successful completion and non-completion. In addition, NN has a higher performance that can be attributed to the flexibility of the technique to model nonlinear relationships and being able to work with nonparametric models.

Table 5. Sensitivity analysis summary of average error

Variable	Average error
FT	116.2703
1st Generation	70.5105
Degree	42.8871
College GPA	22.9242
Plans to work	17.0291
Gender	13.5365
Age	11.088
ACT reading	4.8435
ACT comp	2.7995
High school GPA	2.7799
ACT English	2.1128
ACT math	1.7227

Table 6. Comparison of model performance

	Correct classification rate		Overall Performance
	Complete	Non-complete	
Logistic regression model	98%	91%	81.50%
NN model	96.32%	95.16%	88.10%

5. CONCLUSIONS AND FUTURE WORK

This research presented the application of NN for forecasting program completion of community college students in STEM majors with high performance in prediction.

Also, it can be used for identifying the level of importance of the factors impacting such prediction.

Based on this study, the factors impacting prediction of student success, specifically in STEM majors, are consistent with prior research, suggesting that the level of importance of those factors does not depend on the methodology used. However, further research is required to determine if other methodologies imply the same.

Future research should also investigate other factors to determine with more recall the set of factors that impact completion rates among community college students. As a prediction model with the right set of variables can provide a useful tool for the creation of retention strategies by addressing advising strategies.

During the study, several limitations were considered. One limitation of the current study was that the dataset provided from the institution did not include socioeconomic data, which can have an interesting impact in the generation of strategies for retention and student success as stated in prior literature. To assess this limitation further work can be done widening the data collection in number of examples and variables to be included (socioeconomical aspects). Further, data was only considered from one educational institution. Additional studies should be conducted on other universities and using multiple universities.

REFERENCES

Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. Journal of Applied Research in Higher Education, 10(1), 61-75.

Alkhasawneh, R., & Hargraves, R. H. (2014). Developing a hybrid model to predict student first year retention in STEM disciplines using machine learning techniques. Journal of STEM Education: Innovations and Research, 15(3), 35-42.

Babić, I. D. (2017). Machine learning methods in predicting the student academic motivation. Croatian Operational Research Review. 8, 443-461.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. Decision Support Systems, 49(4), 498-506.

Delen, D. (2011). Predicting student attrition with data mining methods. Journal of College Student Retention: Research, Theory & Practice, 13(1), 17-35.

Dissanayake, H., Robinson, D., & Al-Azzam, O. (2016, January). Predictive Modeling for Student Retention at St. Cloud State University. In Proceedings of the International Conference on Data Mining (DMIN) (p. 215). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Hassoun, M. H. (1995). Fundamentals of artificial neural networks. MIT press. Cambridge, MA.

Haykin, S. S. (2009). Neural networks and learning machines/Simon Haykin. New York: Prentice Hall.

Miranda, M. A., & Guzmán, J. (2017). Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos. Formación universitaria, 10(3), 61-68.

Morris, L. V. (2016). Mining Data for Student Success. Innovative Higher Education, 41(3), 183-185.

Oztekin, A. (2016). A hybrid data analytic approach to predict college graduation status and its determinative factors. Industrial Management & Data Systems, 116(8), 1678-1699.

Radunzel, J., & Noble, J. (2012). Predicting Long-Term College Success through Degree Completion Using ACT [R] Composite Score, ACT Benchmarks, and High School Grade Point Average. ACT Research Report Series, 2012 (5). ACT, Inc.

Schmitt, N., Keeney, J., Oswald, F. L., Pleskac, T. J., Billington, A. Q., Sinha, R., & Zorzie, M. (2009). Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. Journal of Applied Psychology, 94(6), 1479

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

Snyder, J., & Cudney, E. A. (2018, June), A Retention Model for Community College STEM Students Paper presented at 2018 ASEE Annual Conference & Exposition, Salt Lake City, Utah. https://peer.asee.org/29719

Undergraduate Retention and Graduation Rates. The condition of education (2018). Retrieved December 16, 2018, from https://nces.ed.gov/programs/coe/indicator_ctr.asp

Williford, A. M., & Schaller, J. Y. (2005, May). All retention all the time: How institutional research can synthesize information and influence retention practices. In Proceedings of the 45th Annual Forum of the Association for Institutional Research

Zhang, G., Anderson, T. J., Ohland, M. W., & Thorndyke, B. R. (2004). Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study. Journal of Engineering education, 93(4), 313-320.

V. PREDICTING STUDENT RETENTION USING SUPPORT VECTOR MACHINES

Tatiana A. Cardona[a], Elizabeth A. Cudney[a]

[a]Deparment of Engineering Management and Systems Engineering, Missouri University of Science and Technology

ABSTRACT

Universities and colleges have a constant focus on improving student retention and degree completion rates. Degree completion impacts the reputation of the institution, as it reflects institutional ability to prepare graduates with the specific skills that contribute to society through STEM majors. Colleges and universities collect considerable amounts of student data; however, efforts need to be made to utilize the data to increase student success. For instance, by determining the factors that influence student retention and completion rates, it is possible to improve advising through intentional student advising. To this end, this research presents the application of support vector machines (SVM) to predict degree completion within three years by STEM community college students. SVM enables the classification of the input variables into expected classes, completion and not completion, by maximizing the margin between the points from the different classes constraining the misclassification. The model was developed using data on 282 students with 9 variables. The variables included age, gender, degree, and college GPA, among others. The model results, which include prediction and variables ranking, offer an important understanding about how to develop a more efficient and responsive system to support students.

Keywords: Student retention, support vector machines, degree completion, engineering, education.

1. INTRODUCTION

Colleges and universities collect vigorous amounts of data from students from as soon as they apply to the institution. The improvement in processing that data is vital to obtain positive gains about the factors that influence degree completion rates. To this end, the prediction or forecasting of program completion by the student gives insight into areas in need of development to improve advising (Williford and Schaller, 2005) and increase retention and graduation rates. As defined by the National Center for Education Statistics (NCES), retention rates refer to the proportion of students returning to the same institution the following fall, while graduation rates are students who complete the programs in certain amount of time. These terms were adopted for the development of the present study.

The U.S. Department of Education set a goal of preparing a society with individuals that can “understand, explore and engage with the world”, which are specific skills that can be achieved through STEM majors. However, retention rates for college and university students pursuing STEM majors are low, 69% and 48%, respectively according to Snyder and Cudney (2018). Thus, the literature indicates a critical need to increase student retention in higher education as it reflects institutional commitment to students as stated by Slim et al. (2014) and Morris (2016). Therefore, predicting student retention provides insight into opportunities for intentional student advising, better planning, and development of retention strategies based on student needs (Pereira and Zambrano, 2017).

Recently, machine learning techniques have been applied to analyze educational data focused on retention and graduation rates (Pereira and Zambrano, 2017). Within the models in the current literature, SVM, neural networks (NN), and decision trees (DT) have proven to have superior performance than other machine learning techniques based on prediction accuracy.

The purpose of this research was to develop a prediction model using the SVM technique to forecast program completion within 3 years by STEM students in a Midwest community college. The following research question is investigated:

Can SVM model accurately forecast students at risk of dropout for students in a Midwest community college, specifically, in STEM majors?

Therefore, this research is focused on information processing in order to make the available data useful (Snyder and Cudney, 2017). Further, the goal was to identify the factors that influence successful completion. To this end, this paper presents the application of SVM as a machine learning technique using a database comprised of 282 entries with 9 variables collected from a community college in the Midwest. The remainder of this paper is structured into the following sections: literature review and background on SVM applications on student success prediction, data analysis and predictive model development and validation, and comparison to prior research.

2. LITERATURE REVIEW

SVM enables the classification of the input variables into expected classes, by creating a hyperplane in between and then maximizing the margin between the points from the different classes and the hyperplane to constraint the misclassification (Haykin, 2009). The algorithm can be used in linear and nonlinear models (Suthaharan, 2016). Within the literature, SVM has been one of the most frequently applied machine learning techniques for prediction of student success. Also, SVM had presented high performance when predicting student success, with model accuracy over 77% in all the cases (Delen 2010, McAleer and Szakas 2010, Oztekin 2016). For instance, Delen (2010) used data mining methods such as NN, DT, and SVM to predict student attrition prior to sophomore year. The best results were from the SVM technique with 81.18% accuracy. In McAleer and Szakas (2010), the methodologies used to predict retention risk from past data and determine if transfer students have a higher retention risk were Naïve Bayesian and SVM. SVM obtained a 79.59% performance, which surpassed the results of the Naïve Bayesian model (57.35%). The research also concluded that transfer students do not have increased retention risk. Further, Oztekin (2016) used DT, artificial neural network (ANN), and SVM for the prediction of undergraduate degree completion at a four-year university. The three methods were effective in predicting degree completion, with rates over 70%. The more consistent and highest evaluation rates were found for the SVM model.

The literature has shown that different methodologies have different performance results depending the source of information. SVM had obtained high accuracy when predicting student success; however, in other studies such as Babić (2017), no difference was found between the performance obtained using the three methods when applying a test of significance. The methodology includes the application of machine learning classifiers such as NN, DT, and SVM. All methods had performance rates below 73%.

The literature also illustrated the importance of data preparation in the application of machine learning techniques for unbalanced datasets. For example, Delen (2010) and Delen (2011) found that that machine learning techniques, specifically NN and support SVM, have better performance when working with a balanced dataset. Other studies also undergo a cleaning and balanced process before applying machine learning techniques (Kondo et al. 2017 and Adejo and Connolly 2018).

3. RESEARCH METHODOLOGY

The research process was conducted in the following stages: 1) data description and preparation, 2) data modeling, application of SVM, and 3) model assessment. A pictorial representation of the modeling process is presented in Figure 1. The stages are explained in more detail in the following subsections.

3.1. DATA DESCRIPTION AND PREPARATION

The dataset was comprised of five years of registered student data, which contained 904 students, pursuing degrees in chemistry, biology, or engineering. From this data, 177 were identified as completing the degree within three years, which is 150% of normal time for completion as required to be reported by the 1990 Student Right-to-

Know Act for postsecondary institutions. The remaining 727 students did not complete their degree within that period with reasons considered as of college withdrawal or switching to a non-STEM major. The data set was cleaned due to considerable missing data and inconsistent data. For example, standardized exam scores were missing for some students as this information is not required for community college admission. After cleaning the data to remove incomplete records, the final dataset consisted of 282 students, 131 non-completers and 51 completers. For the present study, reducing the number of variables was necessary for specificity and to avoid redundancy. The number of variables resulting after cleaning the data was moderate for developing the network. The input variables are presented in Table 1 and Table 2.

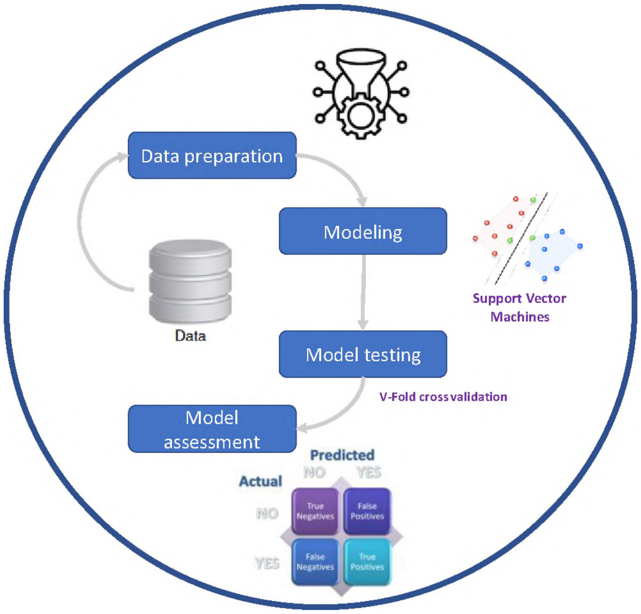
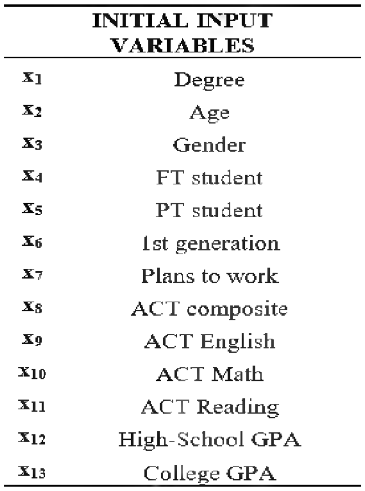


Figure 1. Methodology

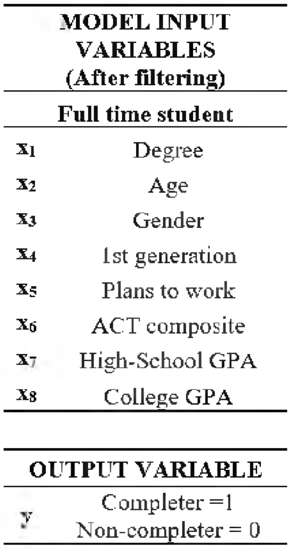
Initial experiments showed high performance in the prediction model but low recall, which could be attributed to the unbalance in the data for the target value (more non-completers than completers). Therefore, it was necessary generate a subsample to balance the number of instances for both classes. Data distribution is Figure 2 initial data (left side) versus balanced data (right side). This was performed using the stratified sampling function in STATISTICA 12.

Table 1. Initial input variables



	INITIAL INPUT VARIABLES
x_1	Degree
x_2	Age
x_3	Gender
x_4	FT student
x_5	PT student
x_6	1st generation
x_7	Plans to work
x_8	ACT composite
x_9	ACT English
x_{10}	ACT Math
x_{11}	ACT Reading
x_{12}	High-School GPA
x_{13}	College GPA

Table 2. Variables used in the model



	MODEL INPUT VARIABLES (After filtering)
	Full time student
x_1	Degree
x_2	Age
x_3	Gender
x_4	1st generation
x_5	Plans to work
x_6	ACT composite
x_7	High-School GPA
x_8	College GPA

	OUTPUT VARIABLE

y	Completer =1
	Non-completer = 0

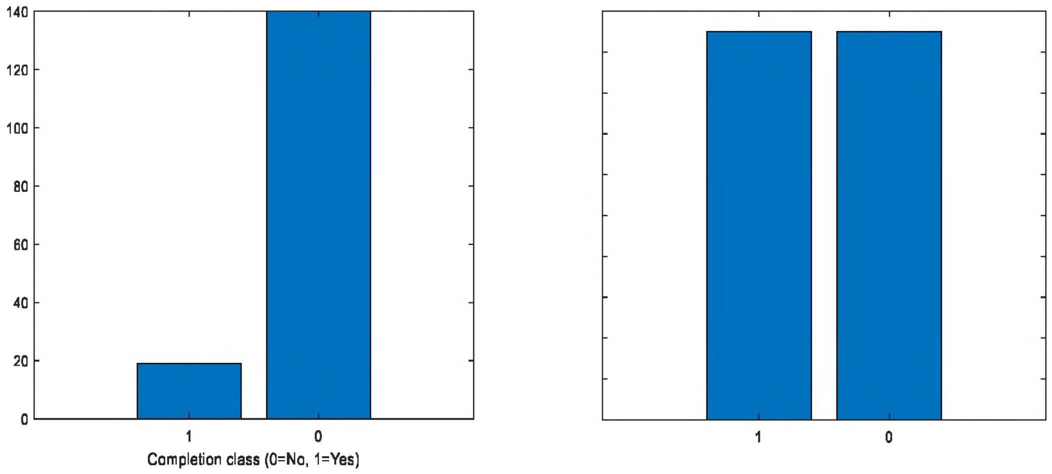


Figure 2. Initial data distribution (left side) vs balanced data distribution (right side)

3.2. MODEL

SVM is a supervised learning algorithm that can perform classification or regression for categorical and numerical response variable, respectively. It creates a mapping space to separate the input data in different classes. The model is capable of mapping linear and non-linear data by deploying kernel functions that can transform the inputs to a higher dimensional space, which allows for a linear separability. Then, the use of kernels reduces the complexity of the problem by creating parallel hyperplanes that separate the data. The optimum condition is found by minimizing the Euclidean norm of the weight vector, which is a constrained optimization problem that can be solved using the method of LaGrange multipliers. The algorithm maximizes the margin between the parallel hyperplanes constraining the misclassification. It is assumed that as the distance increases between the hyperplanes, the generalization error decreases. One of the advantages of using SVM is that it works well with small sample data (Shawe-Taylor and Cristianini 2000), which is the case in the present research.

The model selected was SVM type 2 classification. This model classifies binary data for a discrete target variable. The algorithm used in the classifier was radial basis function (RBF), which can be identified as the kernel for dimensional transformation. The model was implemented using STATISTICA 12.

k-fold cross validation was used for training testing and validating the prediction model. An error goal of 0.01, and a maximum number of iterations of 10,000 were set as stopping criteria. A summary of the model specifications is shown in Table 3.

Table 3. Model summary

Model specifications	Value
No. of independent variables	8
SVM type	Classification type 2
Kernel Type	Radial Basis Function
Number of SVs	82 (26 bounded)
Number of SVs (0)	34
Number of SVs (1)	48

3.3. MODEL ASSESSMENT

The model was assessed using precision and recall measures in the validation set and overall accuracy for the model. The last step ensures a more holistic analysis of the results by mitigating possible misinterpretations. It is important for the model to be precise at predicting non-completers (low error type II) as the results are intended to improve and develop retention strategies, which incur costs for the institution when investing in students that are a false negative for completion risk. The overall performance was calculated as the proportion of correctly classified values from the training, testing, and validation subsamples obtained from the k-fold cross validation application.

4. RESULTS

The summary of the results presented in Table 4 indicate that 26 of 82 vectors were classified as bounded. Bounded vectors are located within the margin area as the model used soft boundaries. These represent only 9% of the classified vectors which give an insight of a good implementation of the model as data generalization is better when the number of bounded vectors is low in proportion of the total examples (Bottou and Lin, 2007).

The best performance of the model was achieved with an error of 0.01 at epoch 2919. Meaning the model achieved the error goal and stopped training. The classification performance (Table 4) recall (false positive) indicated that the model can classify with accuracy over 70% with moderate misclassification. Further, the model is more precise when predicting non-completers. Although no weights were used to prioritize class classification, the results are more accurate for predicting students at risk of dropout (non-completers). This is important to consider when creating retention strategies that are focused on intentional advising, as treating false positive misclassifications can incur some unnecessary cost. This is the reason why the model analysis is focused on the recall measure.

Table 4. Confusion matrix, precision, and recall measures

Class	0	1	Total	Recall
0	39	8	47	0.8298
1	7	17	24	0.7083
Total	46	25	71	
Precision	0.8478	0.6800		

The overall accuracy of the model is high as presented in Table 5. However, there is an evident difference between training and testing performances. In this case, the testing accuracy offers more information about the prediction performance as it prevents misinterpretations related to data overfitting. Then, it can be said the model offers a good prediction performance when testing accuracy is over 78%, which is an adequate measure for the prediction purposes stated in the problem.

Table 5. Model accuracy

Classification accuracy (%)	
Train	94.313
Test	78.873
Overall	90.42

5. CONCLUSIONS AND FUTURE WORK

This research presented a complete case of the application of SVM in predicting degree completion. The model results showed a good performance with recall rates over 70% and testing rates over 78%. Thus, SVM technique provides a good resource for the prediction of student success in a Midwest community college for students in STEM majors. Further, this case study contributes to create evidence of the application of models specifically to community college data, as most of previous literature of machine learning applications for student success is focused on data collected from universities.

Based on the performance of the model, it is possible to determine the variables that have an impact on predicting student success; however, further work is recommended to identify the ranking of impact of each one. The identification of the impact of factors included in the model is of benefit to improve and create more efficient and customized retention strategies.

Some limitations were present during the development of the described model. First, the dataset was not collected specifically for the current research. The number of variables and data points had to be reduced to generate a more adequate sample. This increased the risk of overfitting the model; thus, several combinations of the initial model parameters where tested to determine the most adequate combination.

As future work, the present research could be complemented by extending the model to identify the rank of importance of the variables. In addition, datasets from different institutions could provide further insight of general behavior of completion specifically in community colleges including other factors such as aspects as funding status and demographical characteristics. Further research should also examine other prediction techniques to develop a prediction model for community college students.

REFERENCES

Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. Journal of Applied Research in Higher Education, 10(1), 61-75.

Babić, I. D. (2017). Machine learning methods in predicting the student academic motivation. Croatian Operational Research Review. 8, 443-461.

Bottou, L., & Lin, C. J. (2007). Support vector machine solvers. Large scale kernel machines, 3(1), 301-320

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. Decision Support Systems, 49(4), 498-506.

Delen, D. (2011). Predicting student attrition with data mining methods. Journal of College Student Retention: Research, Theory & Practice, 13(1), 17-35.

Haykin, S. S. (2009). Neural networks and learning machines/Simon Haykin. New York: Prentice Hall

Kondo, N., Okubo, M., & Hatanaka, T. (2017, July). Early Detection of At-Risk Students Using Machine Learning Based on LMS Log Data. In Advanced Applied Informatics (IIAI-AAI), 2017 6th IIAI International Congress on (pp. 198-201). IEEE.

McAleer, B., & Szakas, J. S. (2010). Myth Busting: Using Data Mining to Refute Link between Transfer Students and Retention Risk. Information Systems Education Journal, 8(19), n19.

Morris, L. V. (2016). Mining Data for Student Success. Innovative Higher Education, 41(3), 183-185.

Oztekin, A. (2016). A hybrid data analytic approach to predict college graduation status and its determinative factors. Industrial Management & Data Systems, 116(8), 1678-1699.

Pereira, R. T., & Zambrano, J. C. (2017, December). Application of Decision Trees for Detection of Student Dropout Profiles. In Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on (pp. 528 -531). IEEE.

Shawe-Taylor, J., & Cristianini, N. (2000). An introduction to support vector machines and other kernel-based learning methods (Vol. 204). Cambridge: Cambridge University Press.

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

Snyder, J., & Cudney, E. A. (2017). Retention Models for STEM Majors and Alignment to Community Colleges: A Review of the Literature. Journal of STEM Education: Innovations and Research, 18(3), 48-57.

Snyder, J., & Cudney, E. A. (2018, June), A Retention Model for Community College STEM Students Paper presented at 2018 ASEE Annual Conference & Exposition, Salt Lake City, Utah. https://peer.asee.org/29719

Suthaharan, S. (2016). Machine learning models and algorithms for big data classification. In Integrated Series in Information Systems (Vol. 36). Springer US, Boston, MA.

Williford, A. M., & Schaller, J. Y. (2005, May). All retention all the time: How institutional research can synthesize information and influence retention practices. In Proceedings of the 45th Annual Forum of the Association for Institutional Research.

VI. PREDICTING STUDENT DEGREE COMPLETION USING RANDOM FOREST

Tatiana A. Cardona[a], Elizabeth A. Cudney[a], Jennifer Snyder[b], and Roger W. Hoerl[c]

[a]Deparment of Engineering Management and Systems Engineering, Missouri University of Science and Technology

[b]School of Science, Valencia College's East campus.

[c]Union College, Schenectady, NY.

ABSTRACT

Universities and colleges have a constant focus on improving student retention and degree completion rates. Degree completion impacts the reputation of the institution, as it reflects institutional ability to prepare graduates with the specific skills that contribute to society through STEM majors. Colleges and universities collect considerable amounts of student data; however, efforts need to be made to utilize the data to increase student success. For instance, by determining the factors that influence student retention and completion rates, it is possible to improve advising through intentional student advising. To this end, this research presents the application of support vector machines (SVM) to predict degree completion within three years by STEM community college students. SVM enables the classification of the input variables into expected classes, completion and not completion, by maximizing the margin between the points from the different classes constraining the misclassification. The model was developed using data on 282 students with 9 variables. The variables included age, gender, degree, and college GPA, among others. The model results, which include prediction and variables ranking, offer an important understanding about how to develop a more efficient and responsive system to support students.

Keywords: Student retention, support vector machines, degree completion, engineering, education.

1. INTRODUCTION

Research indicates there is a skill gap in our workforce that will only continue to widen without corrective action in higher education. At the same time, reports indicate that 40 percent of college freshman will not graduate. Therefore, increasing student retention rates in higher education is critical. Also, the ability of these institutions to prepare and graduate students with specific skills is an indicator of institutional performance, making it one of the focus areas for universities and colleges (Williford and Schaller, 2015). This is perhaps more important to community colleges as they are a growing entry point for higher education (Snyder and Cudney 2017). In terms of retention improvement, efforts have been made to adjust admission requirements; however, the retention rates remain low with a national average of 62% for four-year colleges and 60% for universities (Snyder and Cudney, 2018) and many of these strategies have reduced access from different economic sectors to higher education (Kirp, 2019). Thus, many institutions have recognized the need to understand the factors that contribute to retention to better focus their efforts.

While universities and colleges collect considerable student data, their ability to process the available information does not occur at the same pace as the collection (Morris, 2016). There needs to be a method allowing for data utilization and timely implementation to improve student retention. For instance, the creation of predictive models that allow for the recognition of students at risk for attrition will enable timely interventions. By identifying the factors through a prediction model, universities and college can provide intentional student advising and planning. Further, higher education institutions can develop retention strategies that focus on identified student needs that meet their specific campus needs (Slim et al., 2014).

According to the literature, machine learning techniques have been applied to predict student success with high confidence (Cardona et al. 2019). Delen, 2010, conducted several studies to compare methodologies such as neural networks (NN), support vector machines (SVM), decision trees (DT), and random forests (RF), among others. The results indicated that these machine learning techniques had better prediction results than other statistical techniques such as logistic regression (LR) and discriminant analysis.

The purpose of this research was to develop a prediction model using the RF technique to predict student success by science, technology, engineering, and mathematics (STEM) students in a Midwest community college. RF was selected for three main reasons: 1. RF has consistently performed at or near the top of machine learning modeling approaches in a wide range of applications, similar to multilayer NN (i.e., deep learning) according to James et al. 2017. 2. RF also provides insight into the contributions of specific variables to the accuracy of the final model, something that is lacking with most machine learning approaches. 3. The RF algorithm is very stable computationally, more so than NN or SVM, for example.

The time considered for successful degree completion was 150% of normal time for completion. This time was employed for the study in order to be consistent with the 1990 Student Right-to-Know Act, which requires postsecondary institutions to report the rate of students graduating in 150% of the time the program was designed (NCES, 2018). As the data was from a community college, student success was measured as student completion within three years. A student pursuing an associate's degree should complete the degree program in two years. Therefore, a student is considered successful if they complete the program in three years or less.

The following research question was investigated in this study: Does the RF technique, based in its classification accuracy, provide a good resource for the prediction of student success at the Midwest community college for students in STEM majors? If so, what variables that have a higher impact in the prediction of student success? The remainder of this paper is structured as follows. First, a literature review provides background on RF applications for student success prediction. The research methodology is described next. The results of the model are then analyzed and discussed. Finally, the conclusions, research limitations, and future work are presented.

2. LITERATURE REVIEW

Most of the literature on the application of machine learning techniques in education focuses on the use of an individual machine learning technique. Ensemble machine learning techniques combine several machine learning techniques and are commonly used to improve prediction models. However, the number of studies in the literature that use ensemble machine learning techniques such as RF, Boosted Trees (BT), and stacking of other techniques is low with only four journal papers published from 2010 to 2017. The results of ensemble machine learning show consistently high overall classification accuracy that ranges between 79.36% and 81.67%. Thus, it is important to develop models that can nurture the body of knowledge on how ensemble machine learning techniques can improve current models. Research by [8] focused on prediction models for retention prior to sophomore year. The study applied classification methods such as NN, DT specifically the C5 algorithm, SVM, and LR. The results were compared to the use of different ensembles including RF, BT, and information fusion, which stack different predictors. The dataset for analysis was comprised of 16,066 students enrolled as freshmen during 2004 and 2008. A well-balanced dataset was developed such that the classes to predict dropout were equally represented. When using the ensemble with the well-balanced data set, the accuracy of the predictions improved to approximately 80%, which was higher than using the standing alone techniques of SVM and DT. A sensitivity analysis showed the variables that impact at-risk student prediction for this study were student scholarships, loans, and fall GPA. A comparison of models was proposed by Dissanayake et al., 2016 to predict student retention at St. Cloud State University. Principal component analysis (PCA) was used to select linear combinations of the variables that were not correlated with one another. Then, the original database and database after applying PCA were used to compare performance. The study applied six prediction models: k-nearest neighbor (KNN), DT, RF, LR, NN, and Bayesian Belief Networks (BBN). The results showed that the models using the PCA filtered dataset yielded better results. For example, the RF technique presented improvement in all evaluation factors and, together with LR, had the highest accuracy results of 84.77% and 83.07%, respectively. (Sweeney et al. 2016) considered the importance of predicting students' grades in the courses they will enroll in during the next semester. The methodology employed factorization machines (FM), which is an adaptation of second order polynomial regression, along with other regression techniques such as RF, stochastic gradient Descent regression (SGD), KNN, and personalized multiple linear regression (PMLP). The model was used with information for each student or course. The dataset was collected during five years from George Mason University, with a total of 15 terms including summer terms. The model results indicate that PMLP had the lowest error from the individual techniques; however, RF provided more accurate predictions when the data lacked prior student information (i.e., first semester or cold start students). Machine learning techniques were employed by Kondo et al., 2017 to predict at-risk students. The dataset used was obtained from the learning management system (LMS) during the first semester of 2015, which was comprised of records from 202 students. The methodology consisted of using LR, SVM, and RF to predict GPA. Classes for prediction were defined as a 1 if their GPA was greater than the average minus one standard deviation and 0 otherwise, meaning the student was at risk. The models were evaluated on the weekly change of the comparative importance of explanatory variables. Prediction from RF showed more stable behavior in terms of precision and sensitivity. With the weekly analysis, the model was able to identify a ranking of important variables depending on the point in time (i.e., number of weeks after the semester started) that was analyzed.

3. RESEARCH METHODOLOGY

The research process was conducted according to the main steps of data mining, which include the collection of the data to the reporting and use of it (Feelders et al., 2000). Although the data utilized in this study was not specifically collected for the purpose of predicting retention, the data mining steps were applied as represented in Figure 1. The research process is presented in the following segments: 1) data description and preparation, 2) data modeling and application of RF, and 3) model assessment.

3.1. DATA DESCRIPTION AND PREPARATION

The data for this research was collected from a community college in the Midwest that offers associate degrees in STEM majors. The dataset provided by the institution was comprised of 904 students pursuing degrees in chemistry, biology, and engineering. The data collected included information on students registered from spring 2013 through fall 2017. The raw dataset contained a considerable amount of missing and inconsistent data. The reason behind this is that the institution is an open-admission institution; thus, information such as high school GPA and standardized exam scores are not required for admission. Therefore, it was reasonable to remove students that did not report high school GPA and standardized exam scores, as the missing information would highly impact the application of the classifier algorithm for predicting student success. Also, cases with inaccurately reported data (for example, scores out of the normal score range) were not taken in account. Table 1 presents a summary of the descriptive statistics for the numerical variables in the initial dataset. Table 2 shows the variables used in the study.

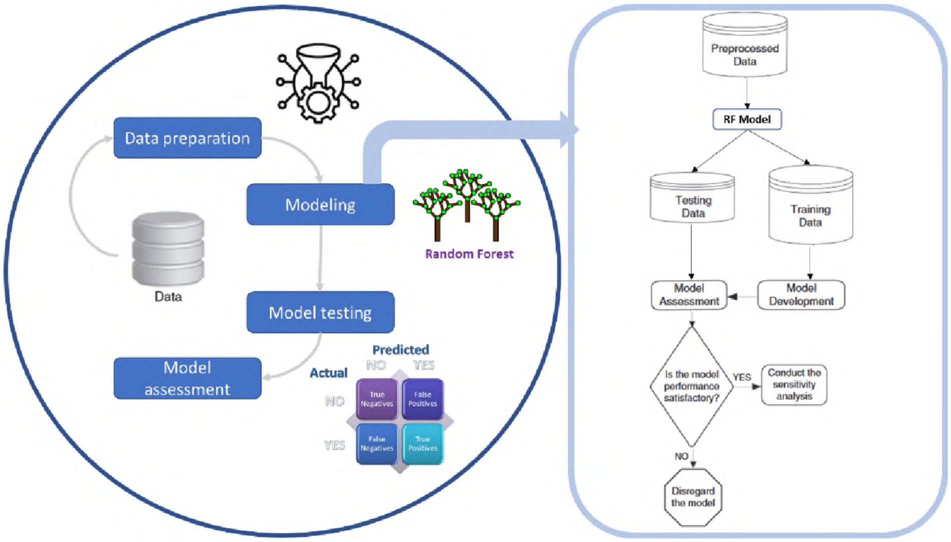


Figure 1. Data analytic methodology

Table 1. Raw data descriptive statistics for numerical variables

Variable	N	Mean	Median	Min	Max
Age	904	24.85	21	16	65
ACT Comp	428	22.64	22	11	34
ACT English	436	22.01	21	7	35
ACT Math	436	22.835	22	13	35
ACT Reading	435	23.13	22	9	36
High School GPA	605	4.13	3.51	1	91.38
College GPA	814	2.775	2.95	0	4.93

Table 2. Variables used in the study

Variable	Type
Complete (Target variable)	Yes/No
Degree	Chemistry, Biology, Engineering
Age	Numerical
Gender	Female/Male
Full Time Student	Yes/No
1st Generation Student	Yes/No
Plan to work	Yes/No
ACT comprehensive	Numerical
ACT English	Numerical
ACT mathematics	Numerical
ACT reading	Numerical
High school GPA	Numerical
College GPA	Numerical

Removing the incomplete records resulted in a final dataset of 282 students, which consisted of 51 completers and 231 non-completers. For this research, completers were defined as the students that completed their associate's degree in three years or less. Conversely, non-completers did not finish their associate's degree within three years. The resulting dataset contained a moderate number of variables (14 variables) for developing the RF model. The input variables are presented in Table 3.2. Variables as age, gender, first generation student, plan to work, high school GPA, and ACT scores were self-reported when the student applied for admission. College GPA was the overall GPA of the student as of fall 2017 or their GPA upon graduation if the student had completed their studies. The degree was the student's current degree as of fall 2017 or their awarded degree if the student had graduated. Initial experiments suggested that it was beneficial to generate a subsample to balance the number of instances of the prediction classes (i.e., completers and non-completers). The initial results provided high overall classification accuracy but low precision (correct predictions out of total predictions of the class). 126 These results are consistent with other studies such as He and Garcia, 2009. Their research focused on imbalanced data and identified several reasons why learning algorithms work better with balanced data. For example, for the DT algorithm the findings indicated that successive partitioning left even fewer examples of the minority class, which reduces the confidence estimates. In addition, the sparseness can blurry characteristics that may result in reducing classification performance.

As RF is a collection of DT, they are sensitive to imbalanced data (Chen and Bermian, 2004). Therefore, the initial performance results in the experimental phase of this study were attributed to the imbalanced data as there were more non-completers

(231) compared to completers (51) as shown in Figure 2, where 1 indicates completion in three years or less and 0 indicates the student did not complete the program in three years or less. Then, a balanced subsample to continue the modeling process was generated using the stratified sampling function in STATISTICA 12 that allowed a user-defined proportion of the minority class to be over sampled in this specific case. Random under sampling and oversampling techniques to balance datasets has been widely used and have shown to improve classifier accuracy [Delen 2010, He and Garcia 2009 and Millar and Richardson 2015).

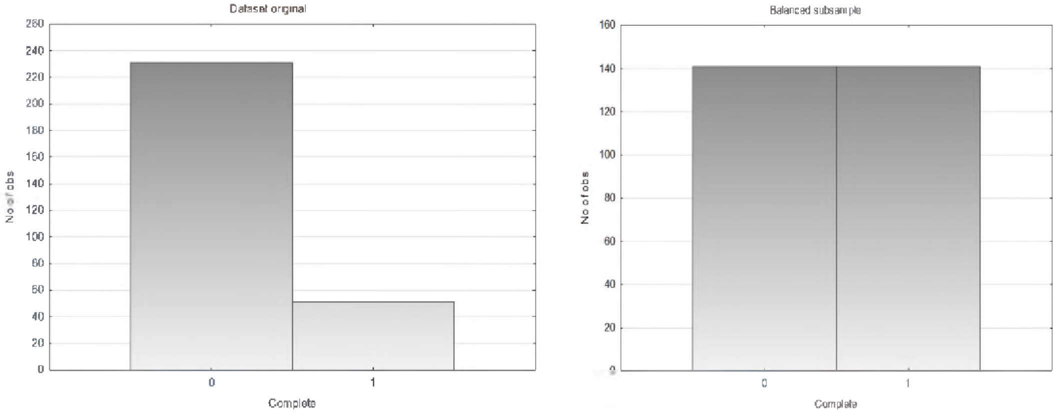


Figure 2. Initial data distribution (left side) vs balanced data distribution (right side)

3.2. DATA MODELING

The RF algorithm is an ensemble of decision trees created randomly from a given dataset. Each tree is created with a different data set chosen randomly (with replacement) from the original data set, a technique known as "bootstrapping." Then, at each branch of each tree, a subset of variables is chosen randomly, and the tree is forced to select from this subset of variables. The intent of this approach is to force the model to consider other variables, besides the most dominant, which might provide greater predictive power with the new data set. The final tree produces a classification response (class prediction) for each observation. This approach is then replicated for numerous trees, producing a “forest.” Each tree generates a vote that enables the classification of the input variables into expected classes, completer and non-completer. The forest then classifies by “majority vote.” The variables that are important for class prediction are also determined based on measures of internal errors (on the tree nodes), tree strength in the forest (classification accuracy), and correlation between the trees. Thus, a more accurate classification is obtained than if analyzing a standing alone DT [18]. Another advantage of this technique is it is not as prone to overfitting as most machine learning algorithms due to the law of large numbers, which states that performing an experiment a large number of times will provide a stable result long term. In other words, the average of the results will be closer to the expected value as more trials are performed. The model was implemented using STATISTICA 12. The parameters used in the training were set as shown in Table 3. Several experiments were run using different combinations of the variable parameters to identify the model with the highest overall classification accuracy. To test the model, a subset comprised of 30% of the original dataset was randomly selected and held until the training was concluded.

3.3. MODEL ASSESSMENT

It is important for the model to be precise at predicting non-completers as the results are intended to improve and develop retention strategies. A retention strategy based on a false negative for completion risk could result in incurred costs for the institution and may not help students. Therefore, the assessment metrics were selected based on the classification accuracy for non-completers precision and recall measures for the testing set and overall classification accuracy for training and testing sets.

Table 3. Modeling parameters

Parameter type	**Parameter**		**Selection**
Fixed	Misclassification cost		Equal
	Prior probabilities		Estimated
	Stopping parameters/each tree:	Max n of nodes	7
		Max n of levels	10
		Min n of cases	7
		Min n in child node	5
Variable	Number of trees		100, 150, and 250
	Model stopping condition:	Percentage decrease in training error (evaluated every 10 cycles)	5%, 1%, and non-stopping condition

The level of importance of the factors that impact the prediction in the model were also identified. Recall that this is a key advantage of RF. STATISTICA calculates the drop in the node impurity and adds the result from every node for each variable. The largest sum represents the most important variable. The ranking score is scaled and presented on a range of 0-100. This measures how often the individual trees split on this variable, and the additional discriminatory power these splits provided.

4. RESULTS

Different parameter combinations were tested including the number of trees with a stopping condition of 5% then with a 1% decrease in error. The results are presented in Figure 3 for the scenarios with 100 and stopping condition of 5% decrease in error (stopped at 70 trees) on the left side and 250 trees with non-stopping condition on the right. As shown in Figure 3, the misclassification for the testing data started to be stable (no significant increase or decrease) after approximately 40 trees. This finding was consistent when using a total of 250 trees. Note that Figure 3 shows both classification accuracy with the original "training" data, used to fit or train the model, and also with test data that was held out from fitting the model.

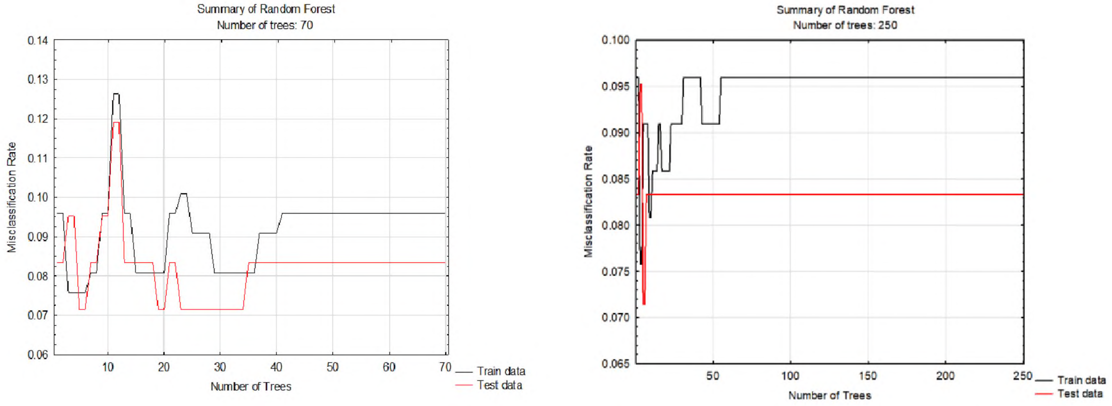


Figure 3. Misclassification rate.70 trees (left), 250 trees (right)

The overall accuracy of the model for the training and test subsets is displayed in Table 4. There is not a significant difference between the overall accuracy performance for the training and testing subset. The results indicate that RF offers a good prediction model for STEM degree completion for the Midwest community college students with a validation performance of approximately 91%. For higher education institutions, this classification accuracy for predicting retention rates supports the development of strategic endeavors to increase student success.

Table 4. Model Accuracy

Subset	Overall accuracy
Train	0.904
Test	0.917

The misclassification ("confusion") matrix is provided in Table 5 and recall and precision measures are presented in Table 6. Both results are indicative of high prediction performance for the classification of non-completers. Specifically, for the test subsample precision (95.2%) and recall (88.9%) shows a risk of misclassification under 11%.

Table 5. Confusion matrix. Training subsample (left). Testing subsample (right)

Training				
		Predicted		
	Class	0	1	Total
Observed	0	79	17	96
	1	2	100	102
	Total	81	117	198

Test				
		Predicted		
	Class	0	1	Total
Observed	0	40	5	45
	1	2	37	39
	Total	42	42	84

Table 6. Assessment measures for training and test subsamples

	Training	Test
Recall	0.975	0.889
Precision	0.823	0.952

After evaluating the classification accuracy of the model, it was important to identify the variables that impact the prediction. The information gain (Gini factor for classification models) is used to define the rank of the variables. Each tree is partitioned by choosing the variable that offers a higher information gain (Chakrabarti et al., 2008). To determine the importance of each variable in the tree, STATISTICA uses the sum of the information gain from the overall nodes to find the variable overall information gain.

The rank of the variables is determined by adding the information gain of each variable for all the trees and, scaling it in such way that the highest value will be 100. When the resulting value is less than or equal to zero, the variable does not impact the model and can be removed. Table 7 and Figure 4 present the rank of importance of the different variables used. The results showed that the most significant variables are age, college GPA, ACT composite, and ACT math. Age is shown as a key variable that can be useful to administrators in predicting completion. Further, of the various academic metrics available, college GPA is the most useful, at least with this data. Although this information could clarify the variable interaction of age with success, as a standalone variable it is not a variable that can govern the student success behavior.

Table 7. Variable importance rank

Variable	**Rank**
Age	100
College GPA	60
ACT Comp	38
ACT Math	33
High School GPA	33
ACT English	32
ACT Reading	27
Degree	17
Part time student	17
Full time student	13
Plans to work	9
Gender	8
First generation	7

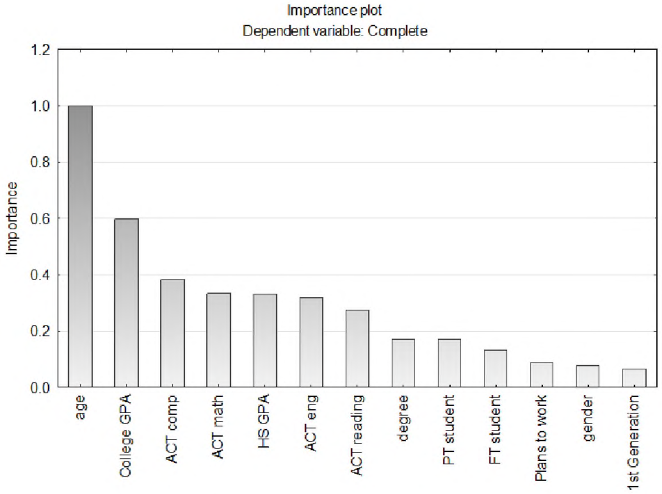


Figure 4. Predictor importance

5. CONCLUSIONS, LIMITATIONS AND FUTURE WORK

This research presented a complete case of the application of RF for predicting degree completion. The model results showed a good performance with precision rates over 80% and testing overall accuracy also over 80%. Therefore, RF technique provides a good resource for the prediction of student success at the Midwest community college for students in STEM majors. Further, this case study contributes in creating evidence of the application of models specifically to community college data, as most of previous literature of machine learning applications for student success is focused on data collected from universities. RF can also be used to identify the level of importance of the factors impacting students successfully completing a degree program. Although GPA is a common factor found in prior literature as important for predicting student success, variables such as ACT math and ACT English are not commonly found as variables of high impact in other studies. In addition, age is also a key variable, which was a similar finding to other studies. Further, the findings suggest that the level of importance of those factors depended on the methodology used; however, further investigation should be performed. Several limitations were present during the development of the described model. In this case, the dataset was not collected specifically for the current research. The number of variables and data points had to be reduced to generate a more adequate sample. This increased the risk of overfitting the model; thus, several combinations of the initial model parameters where tested to determine the most adequate combination. Also, it is important to highlight that, while the study achieved a high classification performance, the data is only representative of one community. Therefore, the results are not generalizable. However, the methodology can be used by other higher education institutions to determine the factors of importance. Further research should be conducted to include other factors such as financial status and other demographic characteristics. This will enable the development of retention strategies and intentional advising that will better address and improve student success. Also, different machine learning techniques should be employed to offer a comparison in performance and a better understanding of the benefits of each approach. Finally, it would also be interesting to analyze the general behavior of student completion for community colleges by collecting information from different institutions. This may help identify factors that vary by institution which may later become retention issues.

REFERENCES

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Cardona, T., Cudney, E. A., & Snyder, J. (2019). *Predicting degree completion through data mining*. Proceedings of the ASEE Annual Conference & Exposition, Tampa, FL.

Chakrabarti, S., Cox, E., Frank, E., Güting, R. H., Han, J., Jiang, X., ... & Pyle, D. (2008). *Data mining: know it all*. Morgan Kaufmann.

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*, *110*(1-12), 24.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498-506.

Dissanayake, H., Robinson, D., & Al-Azzam, O. (2016, January). predictive modeling for student retention at St. Cloud State University. In *Proceedings of the International Conference on Data Mining (DMIN)* (p. 215). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Feelders, A., Daniels, H., & Holsheimer, M. (2000). Methodological and practical aspects of data mining. *Information & Management*, *37*(5), 271-281.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263-1284.

James, G. (2017). D, Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning with Applications in R.

Kirp, D. (2019). The college dropout scandal. The Chronicle Review. https://www.chronicle.com/interactives/20190726-dropout-scandal

Kondo, N., Okubo, M., & Hatanaka, T. (2017, July). Early detection of at-risk students using machine learning based on LMS log data. In *Advanced Applied Informatics (IIAI-AAI), 2017 6th IIAI International Congress on* (pp. 198-201). IEEE.

Millard, K., & Richardson, M. (2015). On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping. *Remote sensing*, *7*(7), 8489-8515.

Morris, L. V. (2016). Mining data for student success. *Innovative Higher Education, 41*(3), 183- 185.

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

Snyder, J., & Cudney, E. A. (2017). Retention models for STEM majors and alignment to community colleges: A review of the literature. *Journal of STEM Education*, 18(3), 30-39.

Snyder, J., & Cudney, E. A. (2018, June), *A retention model for community college STEM students* Paper presented at 2018 ASEE Annual Conference & Exposition, Salt Lake City, Utah. https://peer.asee.org/29719

Sweeney, M., Lester, J., Rangwala, H., & Johri, A. (2016). Next-term student performance prediction: A recommender systems approach. *JEDM| Journal of Educational Data Mining*, *8*(1), 22-51.

Undergraduate Retention and Graduation Rates. The condition of education (2018). Retrieved December 16, 2018, from https://nces.ed.gov/programs/coe/indicator_ctr.asp

Williford, A. M., & Schaller, J. Y. (2005, May). All retention all the time: How institutional research can synthesize information and influence retention practices. In *Proceedings of the 45th Annual Forum of the Association for Institutional Research*

SECTION

2. CONCLUSIONS AND RECOMMENDATIONS

The systematic review presented as Paper I of this document, offered a significant information of the current panorama of the application of machine learning techniques to predict student success. Machine learning techniques have been applied in education to predict retention and identify factors that influence retention rates for several years, with more successful results since 2010. The most frequently used techniques were DT, NN, and SVM. Also, other models such as ensembles have been developed that have shown accurate classifications. It was also found that although novelty models have been developed, they were customized for segments within each institution. Also, the list of factors in the models changed depending on the study. A consistent list of factors that can be scalable to other institutions for prediction of degree completion has not been identified in the literature.

This review leads to conclude that institutions should develop synchronized systems that are able to collect student data that feed the learning algorithms in order to have the most benefit from them. As it is statistically assumed, the more data the more reliable are the results. However, it is also important to highlight from this systematic review that the algorithms have proved to be efficient for predicting student success using less than 68 variables. This means that the studies can be segmented, and specific datasets can lead to specific analysis. As stated by Essa and Ayad (2012) "Decomposition provides a flexible mechanism for building predictive models for application in multiple contexts.” Meaning by decomposing the application of the model in different scenarios of the institutions, more flexible models can be developed.

With further investigation on the factors that impact student success it was possible to propose an architecture for a higher education system for the prediction of student success. The proposed system offers a clear picture of the interaction of the students’ characteristics and their evolution through the course of the college experience. Further, it can be used as a base for formulating models that study student success.

The architecture was validated with promising results for the prediction of student completion for bachelor’s degree data collected from a university in the Midwest of the country. NN was used in the validation and the prediction accuracy obtained was above 98%. This work should be regarded as a preliminary effort to incorporate external and internal factors that impact the success of the students in higher education and how that emergent behavior can be predicted. Also, it continues demonstrating, as in prior literature, that the NN technique is an appropriate tool for student success prediction.

Further, using data from a community college in the Midwest, the system was also validated using several machine learning techniques, including decision trees, neural networks, support vector machines, and random forest. All the techniques showed high classification accuracy in the prediction of student completion (over 80%). Random forest was the best performing technique from those methods with a classification accuracy of 91% for the test subsample. In prior literature, only a few studies use ensembles such as random forest; however, it is not conclusive that they represent a better option for the prediction of student retention. Future research should focus on using ensemble techniques to nurture the body of knowledge on what mixtures of machine learning techniques can provide higher accuracy.

The results of the models elaborated from this system can enable the creation of strategies and reforms. The goal of those strategies and reforms would be increasing student completion rates. Lessons learned will also nurture the body of knowledge of accepted strategies and reforms that can be scaled and applied in other institutions.

One of the limitations to this study is the validation was done using data that was previously collected and readily available. Thus, not all categories of factors proposed in the system were represented. This, reduced the scope of the validation to be only an effort to determine the effectiveness of the architecture

It is also important to note that the proposed methodology can be applied to other institutions. However, the level of impact of the variables used in the prediction is inherent to the institution where the data was collected from. Further analysis of the system could be evaluated by the design of a model that incorporates a more comprehensive set of factors. It is key that institutions study student success models and strategies in a progressive and broader manner, not only for small segments of the system. It should be based on holistic knowledge of how the system impacts students in their career journey. IT should be considered as the starting point to reconcile efforts to improve completion rates and success in general. To generate this support, there must an environment of trust, due to the sensitivity of the information and data that institutions can collect. Therefore, the reforms and strategies that can be formulated should be accompanied with the establishment of policies for evidence-based analytics that encompass the model-data transparency (collection and usage) to legal and ethical clarity.

BIBLIOGRAPHY

Alkhasawneh, R., & Hargraves, R. H. (2014). Developing a hybrid model to predict student first year retention in STEM disciplines using machine learning techniques. Journal of STEM Education: Innovations and Research, 15(3), 35-42.

Bahar Baran, & Eylem Kihç. (2015). Applying The CHAID Algorithm to Analyze How Achievement is Influenced by University Students' Demographics, Study Habits, and Technology Familiarity. Journal of Educational Technology & Society, 18(2), 323-335. Retrieved February 13, 2020, from www.jstor.org/stable/jeductechsoci.18.2.323

Bailey, T., Jenkins, D., & Leinbach, T. (2005). Is Student Success Labeled Institutional Failure? Student Goals and Graduation Rates in the Accountability Debate at Community Colleges. CCRC Working Paper No. 1. Community College Research Center, Columbia University.

Bailey, T. (2017). Community colleges and student success: Models for comprehensive reform. Educause Review, 52(3), 33-42.

Cardona, T., Cudney, E., & Snyder, J. (2019a). Predicting degree completion through data mining. Proceedings of the ASEE Annual Conference & Exposition, Tampa, FL.

Carnevale, A. P., Jayasundera, T., & Gulish, A. (2016). America's Divided Recovery: College Haves and Have-Nots. Georgetown University Center on Education and the Workforce. https://cew.georgetown.edu/wp-content/uploads/Americas-Divided-Recovery-web.pdf

Dahlstrom, E. (2016). Moving the red queen forward: Maturing analytics capabilities in higher education. Educause Review, 51(5), 36-54.

Governor's Business Council, T. (2002). Building An Effective and Aligned P-16 Education System: What Should Higher Education Do to Enhance Student Access and Success?. ERIC Clearinghouse.

Grajek, S. (2017). Top 10 IT Issues: Foundations for Student Success. Educause Review, 52(1).

Handel, S. J. (2013). The transfer moment: The pivotal partnership between community colleges and four-year institutions in securing the nation's college completion agenda. New Directions for Higher Education, 2013(162), 5-15.

Hiles, H. (2017). Student success, Venture Capital, and a Diverse Workforce: An Interview with Heather Hiles. Educause Review, 52(3), 22-30.

Iam-On, N., & Boongoen, T. (2017). Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings. International Journal of Machine Learning and Cybernetics, 8(2), 497-510.

Jenkins, D., & Fink, J. (2015). What we know about transfer. New York, NY: Columbia University, Teachers College, Community College Research Center.

Kelly, A. P., & Schneider, M. (Eds.). (2012). Getting to graduation: The completion agenda in higher education. JHU Press.

Kirp, D. (2019). The college dropout scandal. The Chronicle Review. https://www.chronicle.com/interactives/20190726-dropout-scandal

Klempin, S., & Karp, M. M. (2018). Leadership for transformative change: Lessons from technology-mediated reform in broad-access colleges. The Journal of Higher Education, 89(1), 81-105.

Kondo, N., Okubo, M., & Hatanaka, T. (2017, July). Early detection of at-risk students using machine learning based on LMS log data. In Advanced Applied Informatics (IIAI-AAI), 2017 6th IIAI International Congress on (pp. 198-201). IEEE.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. Expert Systems, 33(1), 107-124.

Matthews, D. (2012). A Stronger Nation through Higher Education: How and Why Americans Must Achieve a Big Goal for College Attainment. A Special Report from Lumina Foundation. Lumina Foundation for Education.

Pereira, R. T., & Zambrano, J. C. (2017, December). Application of decision trees for detection of student dropout profiles. In Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on (pp. 528-531). IEEE

Restuccia, D., & Taska, B. (2018). Different skills, different gaps: Measuring and closing the skills gap. Developing Skills in a Changing World of Work: Concepts, Measurement and Data Applied in Regional and Local Labour Market Monitoring Across Europe, 207. https://www.burning-glass.com/research-project/skills-gap-different-skills-different-gaps/

Shapiro, D., Dundar, A., Huie, F., Wakhungu, P. K., Bhimdiwala, A., & Wilson, S. E. (2018). Completing college: A national view of student completion rates – Fall 2012 cohort (Signature Report 16). Herndon, VA: National Student Clearinghouse Research Center.

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

Thomas, L., Herbert, J., & Teras, M. (2014). A sense of belonging to enhance participation, success and retention in online programs.

Williford, A. M., & Schaller, J. Y. (2005, May). All retention all the time: How institutional research can synthesize information and influence retention practices. In Proceedings of the 45th Annual Forum of the Association for Institutional Research.

VITA

Tatiana Alejandra Cardona Sepulveda completed her Bachelor of Science in Industrial Engineering at Technological University of Pereira, Colombia in 2009. She received her Master of Science in Engineering Management in December 2016 from Missouri University of Science and Technology (S&T). Her research interests included statistical modeling, operations research, and data science. During her PhD studies she received the Omurtag Graduate scholarship for application of systems processes to research problems. Tatiana served as an instructor for four semesters in project management and operations management at S&T. She received a Doctor of Philosophy in Systems Engineering from Missouri S&T in August 2020.