Leveraging machine learning techniques to analyze computing persistence in undergraduate programs

Leila Zahedi, Florida International University

Leila Zahedi is a Ph.D. student in the School of Computing and Information Science (SCIS) at Florida International University. Her main focus is on educational data science and machine learning. Her current research focuses on broadening participation in computing fields in order to attract minorities.

Stephanie J. Lunn, Florida International University

Stephanie J. Lunn is a Ph.D. candidate in the School of Computing and Information Sciences at Florida International University (FIU). Her research interests span the fields of computing education, human computer interaction, data science, and machine learning. Previously, Stephanie received her B.S. and M.S. degrees in Neuroscience from the University of Miami, in addition to a B.S. degree in Computer Science from FIU.

Dr. Samira Pouyanfar, Microsoft

Samira Pouyanfar earned her Ph.D. in Computer Science from Florida International University, Miami, USA in 2019. She received a Master degree in Artificial Intelligence from Sharif University of Technology in 2012 and a Bachelor degree in Computer Engineering from University of Isfahan in 2009. Her research interests include Artificial Intelligence, data science, machine learning, deep learning, and big data. She has published over 30 research papers in international journals and conference proceedings. She is currently working as a data scientist at Microsoft Corporation in Seattle, Washington.

Dr. Monique S Ross, Florida International University

Monique Ross earned a doctoral degree in Engineering Education from Purdue University. She has a Bachelor's degree in Computer Engineering from Elizabethtown College, a Master's degree in Computer Science and Software Engineering from Auburn University, eleven years of experience in industry as a software engineer, and three years as a full-time faculty in the departments of computer science and engineering. Her interests focus on broadening participation in engineering through the exploration of: 1) race, gender, and identity in the engineering workplace; 2) discipline-based education research (with a focus on computer science and computer engineering courses) in order to inform pedagogical practices that garner interest and retain women and minorities in computer-related engineering fields.

Dr. Matthew W. Ohland, Purdue University at West Lafayette

Matthew W. Ohland is Associate Head and Professor of Engineering Education at Purdue University. He has degrees from Swarthmore College, Rensselaer Polytechnic Institute, and the University of Florida. His research on the longitudinal study of engineering students, team assignment, peer evaluation, and active and collaborative teaching methods has been supported by the National Science Foundation and the Sloan Foundation and his team received for the best paper published in the Journal of Engineering Education in 2008, 2011, and 2019 and from the IEEE Transactions on Education in 2011 and 2015. Dr. Ohland is an ABET Program Evaluator for ASEE. He was the 2002–2006 President of Tau Beta Pi and is a Fellow of the ASEE, IEEE, and AAAS.

Leveraging machine learning techniques to analyze persistence in undergraduate computing programs

Abstract

Although student retention remains a significant concern for all Science, Technology, Engineering, and Mathematics (STEM) fields, it is particularly problematic in computing, where enrollment in such programs has not kept pace with the industry demands. Thus, finding meaningful patterns in historical data can help education researchers to reveal the possible reasons for students' withdrawal from a university, and can provide guidelines and mechanisms that lead to improving retention rates. To achieve this goal, we considered the importance of different factors in the graduation of computing students, and generated a predictive model for student graduation using the Multiple-Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD) dataset. We observed that considering input and environment educational variables, cumulative GPA, number of terms registered, start year, institution, and being a transfer student, are the most important features, respectively. Our results also demonstrate that Random Forest algorithm produced a more accurate result on this dataset compared to other machine learning algorithms. We anticipate findings from this ongoing work will give insight to the computing education community and researchers to better understand the relative success of computing students, and that this will, in turn, enable strategic solutions to attain higher retention rates.

Introduction

In general, jobs in STEM fields continue to flourish relative to positions in other fields, however, computing graduates (consisting of students in computer science, computer engineering, and information sciences) are particularly in demand. It is estimated that within a ten year span, growth will increase 32% for information security analysts and 26% for software developers [1]. Despite the professional need for more graduates, undergraduate students in computer or information science have a 59% rate of attrition, which is the highest rate relative to students from other STEM majors [2]. In part, this is due to retention challenges in computer science, although many factors may play a role in a students' decision to drop out. Since pathway patterns for computing students are different from engineering students, it is crucial to specifically explore the variables that contribute to positive academic outcomes in computing fields, to find ways to improve the graduation rate [3].

In order to understand what variables are the most important to student graduation in undergraduate computing programs, we apply Alexander Astin's Input-Environment-Output (I-E-O) model to a filtered version of the Multiple-Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD). Specifically, our version of MIDFIELD

considers a subset solely composed of computing students. We implement Machine Learning (ML) algorithms to predict and to explain the reasons behind computing students' attrition. Then, we leverage the best predictive model to analyze which factors are most critical for graduation in a computing field.

ML is a type of algorithmic and statistical modeling technique that can perform tasks based on pattern recognition and inference, without explicit instructions [4]. We apply several different ML algorithms to assess the following research questions: *1) Which ML algorithm most accurately predicts student outcomes in terms of graduation?*; *2) What indicators predict computing students' graduation?*; and *3) Is there a difference, in terms of importance of the rankings, if the inputs and environment variables are run together, as opposed to when they are run separately?*

The present study is novel in several areas. Although many studies have examined retention and graduation of engineering and computing students, often these studies only consider a short time span of one to two years [5]. Moreover, they tend to focus solely on a single university, leading to results that may not generalize to other students. In this study, we consider approximately 30 years of data for students from 14 universities, to evaluate the influence of individual and institutional factors on computing students' graduation. Having a greater range of dates and institutions from MIDFIELD will provide a more accurate portrayal of the contribution of individual variables. To the best of our knowledge, this is the first time that ML has been applied to MIDFIELD. In addition, since previous research shows that computing majors have different patterns from other STEM majors, the data examined in this paper is exclusive to computing majors [3].

In this document, we will first review the background work in the *Related Research* section. Then, we will discuss framework driving this research in the *Theoretical Framework* section. In the *Database* section, we will cover the dataset that we utilized, as well as what was performed to pre-process and analyze the data. In the *Methods* section, we detail the model evaluation and validation and provide a justification for the final models that we use for our results, as well as providing the actual outcomes in the *Results* section. Moreover, in the *Discussion and Conclusions* section, we provide a discussion of our findings and finish with suggestions for future work in the field.

Related Research

Despite the need for additional graduates in computing fields, the popularity of computing as a major has been declining since the late 1990s [6]. Thus, it is important to understand the factors that contribute to success in computing enrollment, persistence, graduation, and employment. Previous research has demonstrated that demographic characteristics (such as gender and race/ethnicity) [7], and institutional characteristics [8] can affect graduation rates. However, in addition to a lack of popularity in computing fields and industry, there is also a concerning lack of diversity. Relative to the general population, there is a disproportionate amount of White and Asian males in computing [9]. Although numerous studies have begun to examine how to equalize enrollment and retention, levels still remain low.

One technique leverage to better understand these trends is data mining. Data mining has become an important tool to aid in the analysis of educational information for decision making purposes.

By uncovering trends and patterns that exist in repositories, researchers are able to uncover what factors contribute to positive academic outcomes [10–12]. In one study using classification algorithms to examine the contribution of different factors to the graduation rate, it was observed that graduation rates were higher for students that were born and inhabiting the same city. Moreover, they found that transfer students coming from another higher education institution, had increased graduation as well. However, it should be noted that these results may be limited in that they only considered the students of a single private institution, which may affect how well these findings can be generalized [12].

When exploring which factors from high school are most predictive of college graduation, between standardized test scores (SAT and ACT) and students' high school GPA, GPA is consistently considered the winner, in terms of which variable has the greatest impact [13–15]. The hypothesized rationale for this observation is that although standardized tests consider intellectual abilities in certain domains, the overall GPA considers different intrapersonal qualities as well that were useful for positive outcomes in college [15]. More specifically, although grades certainly do reflect skill levels on specific content, it may also include individual factors such as students' attitudes, their behaviors, and the effort exerted.

When specifically considering students from STEM fields, results have shown similar outcomes; however, results vary depending on the measure of academic success and the variables considered. In the realm of engineering, Zhang *et al*. utilized data collected over multiple institutions to apply a multiple logistic regression model to understand the correlation between individual demographic and academic characteristics and graduation [16]. They observed that the greatest predictor of graduation for all the universities considered was high school GPA, and the SAT math scores. Others examining a sample of STEM undergraduates from Georgia Tech, examined the factors that contributed to academic success in college [17]. In this work, academic success included not only graduation rates, but also STEM persistence, and gender differences in grade. Their findings reified that high school GPA is the best predictor of academic success in college, and they also discovered that the next greatest predictor was average score on Advanced Placement examinations. Yet another study examining high school variables - grades in math and science courses, standardized test scores, and high school GPA- for STEM students, found that only performance in high school math and science courses (among which high school calculus was the most important) was predictive of performance in STEM courses in college [18].

Theoretical Framework

The Input-Environment-Output model establishes a framework to evaluate students' effectiveness by linking together the consideration of students' qualities/features upon entry to an educational institution, the impact of their educational environments, and then the students' qualities/features upon exit from the institution [19, 20]. According to this model, *inputs* are considered the characteristics/qualities of a student that exist at their time of enrollment, and may include either static attributes, or those which change with time. Contrarily, *environment* characteristics include the variables that affect a student throughout the course of their educational program. Astin further stratifies these measures as those which are reflective of the overall institution, as well as the specific educational experiences which a student may encounter at the intuition. *Outputs*, are considered the intended goal for the educational program [20].

Table 1: I-E-O Model Depiction of Variables Assessed

Inputs	Environment	Outputs
SAT (Math)	Cumulative GPA	Graduation Rates
SAT (Verbal)	Terms Registered	
ACT	Transfer Status	
Race	Start Year	
U.S. Citizen	Term Entered	
Age	Institution	
Gender	CIP2 (Major)	
	COOP	

We apply the I-E-O model as our overarching theoretical framework, and consider the characteristics described in Table 1 as the specific variables defining our model. Note, GPA stands for Grade Point Average, CIP stands for the students' major during this term, which is expressed as IPEDS (Integrated Post-secondary Education Data System) Classification of Instructional Programs (CIP) code, and COOP (Co-Operative education program) in which a student participates in a partnership between their academic institution and an employer to obtain practical experience through rotations of employment and course study [21]. In addition, the Scholarship Aptitude Test (SAT) and the American College Testing (ACT) are standardized tests used for college admissions in the United States. While the objective of both is similar, the test structure is different, and the SAT includes separate verbal and math scores, whereas the ACT provides a comprehensive score derived from sections on English, Math, Science Reasoning, and an optional essay [22]. In addition, it should be mentioned that although several studies have previously considered High School GPA in addition to, or in lieu of, college GPA- we chose to exclude this measure since different high schools may use different scales, and therefore, the values may be unreliable.

Ultimately, we utilize a reduced dataset specific to computing students, to apply ML algorithms to link learning activities to learning outcomes, which in our context includes retention and graduation. An overview of this approach is illustrated in figure 1. However, it should be clarified that in this model, there are two paths that can be taken in Astin's Model, either: 1) *Inputs* $\overset{A}{\rightarrow}$ *Environment* $\overset{B}{\rightarrow}$ *Outputs*; or directly via 2) *Inputs* $\overset{C}{\rightarrow}$ *Outputs*. Since we do not know if there is a greater impact of $\overset{A}{\rightarrow}$ to $\overset{B}{\rightarrow}$, or to go directly though $\overset{C}{\rightarrow}$, we later assess if there are differences in terms of rankings when we run the inputs and environment variables alone to predict graduation rates, or as a combined set.

Database

To assess what variables are most important to ensuring persistence, we utilized empirical data from the MIDFIELD dataset [21]. MIDFIELD consists of data collected from over 1.5 million undergraduate, degree-seeking students from 19 different institutions. It is a longitudinal student record level database, which means it includes everything that appears on students' transcripts, and that it contains tracking information on each student during their academic career. It considers not only demographic student information (such as sex, ethnicity, and age), but also academic information (such as their major, enrollment status, term and year in which the student

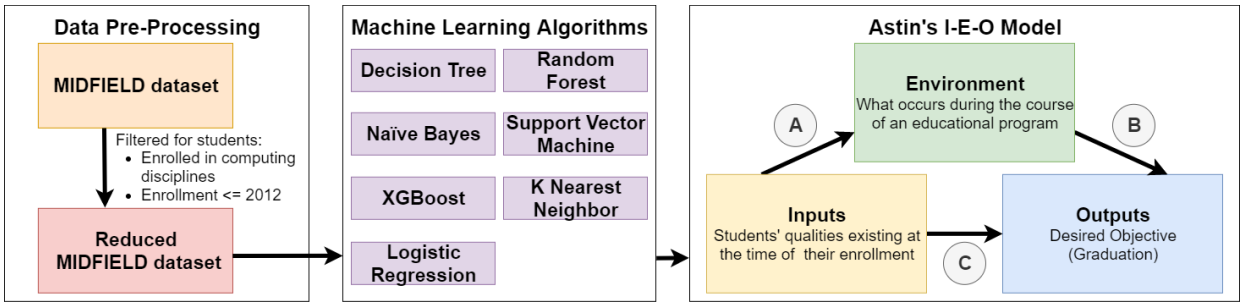


Figure 1: *Adapted version of Astin's I-E-O model applied to computing students to assess persistence/graduation rates. Includes pre-processing of the dataset, and application of several machine learning algorithms.*

graduated, and what type of degree was awarded).

Although MIDFIELD includes students from all engineering disciplines, since we are interested in computing students, we limited our sample to computing majors. The identified students, at some point, were enrolled in one of the following computing disciplines: computer engineering, software engineering, computer science, computer programming, computing and information sciences, and/or information technology. This reduced dataset includes 53,000 students from 14 institutions. Moreover, although MIDFIELD includes data from 1988-2018, we further restricted the dataset to students that enrolled before, or during, 2012. The reason for this pre-processing was that by including students after this timepoint, the data might include students who enrolled, but that had not yet finished their degree- which might skew the sample and make it appear as though less had graduated than actually had. However, the timeframe assessed is consistent with the definition commonly used by the National Center for Educational Statistics, which suggests six years [23]. The pre-processing changes to ensure a subset solely of computing students, resulted in a final reduced set with an N = 39,994, where N is the total number of students in this dataset.

We consider the following variables from the reduced MIDFIELD: SAT math score, SAT verbal score, comprehensive ACT score, race, U.S. citizenship status, age, gender, cumulative GPA, terms registered, term entered, transfer status, start year, institution, CIP2, COOP, and graduation rates. Graduation rate is used as the major outcome measure in our work. However, since it has been demonstrated that the time it takes for a student to obtain a degree may be variable depending on individual and environmental factors (e.g., if they had an external job during their studies, if they studied abroad, etc.) [7, 8], we eliminate any potential bias which may arise from including time as part of the equation. Instead, we consider graduation rate as for all students that finished their degree.

Methods

Having missing values can impact a model's quality when employing machine learning [24]. To impute the data, multiple options exist, ranging from the use of the mean/median values to replacing the absent instances with the most frequent values or zero/constant values. Our dataset was only missing values for SAT and/or ACT scores, since high school students are not required to complete both for college admission, and they typically take either one standardized test or the

Table 2: Measures Considered to Assess Our Algorithms' Performance

Measure	Description of What is Measured	Formula
Accuracy	The performance of the model overall, in terms of how well the model identifies relationships and patterns between the variables. It is obtained by taking the ratio of those observed as predicted correctly out of the total observations	$\frac{T_p+T_n}{T_p+T_n+F_p+F_n}$
Precision	Demonstrates the ability of a classification model to only identify the relevant data points. Shows how accurate the model is for the correct positive predictions out of the total positively predicted observations	$\frac{T_p}{T_p+F_p}$
Recall	The proportion of positive observations which are correct, relative to all the observations in the actual class	$\frac{T_p}{T_p+F_n}$
F1 score	Used for accuracy, it balances precision and recall relative to a specific positive class	$\frac{2(PrecisionxRecall)}{Precision+Recall}$

other [22]. To impute our data, we grouped the data by school and then we used the average for any students missing scores in that school.

Ultimately our goal was to understand what factors are most important in education outcomes for computing students. To better understand what variables are critical for predicting graduation, we utilized ML algorithms. All algorithms run, and their corresponding statistical analyses, were managed using **R** version 3.6.1 in **RStudio**, version 1.1.456.

In order to assess the success of the ML algorithms employed, we considered several evaluation metrics, described further in Table 2. Note that T_p refers to a true positive, T_n to a true negative, F_p to a false positive, and F_n to a false negative. T_p is an outcome in which a model predicts the positive class correctly, and a T_n is an outcome where a model predicts the negative class correctly [24]. Contrarily, a F_p is an outcome in which a model predicts the positive class incorrectly, and in a F_n, the model predicts the negative class incorrectly.

F1 score is considered particularly helpful with uneven class distributions, and considers both false positives and false negatives [25]. Since the cost of our F_p and F_n is similar, F1 score is used the major determinant of the algorithms' performance, along with accuracy. However, we do calculate precision and recall since these provide useful feedback and are required for the estimate of the F1 score.

In machine learning, data modeling can occur using either supervised or unsupervised methods [24]. Unsupervised learning is a technique whereby unlabeled data is applied for a model to discover information on its own. However, supervised learning is a technique using well labeled data, that operates under the assumption that one already knows what the are looking for. We applied the following supervised ML algorithms to analyze our dataset:

- **Decision Tree (DT)**: Works for input/output variables that are continuous and for those that are categorical. However, it is important to note that DTs can be highly sensitive to small changes in the data, resulting in greater variation in the final tree estimation [26, 27].

- **Random Forest (RF)**: Another type of DT learner, it is called a "forest" because many DTs are used and aggregated to produce a class prediction, and is considered "random" because the trees are built differently. In regards to the randomness, each tree uses random samples and random features, so that among the trees, they neither use all the attributes nor all the instances [28–30]. Initially, it uses bagging, an ensemble technique, which combines prediction of individual trees to make overall predictions.
- **Support Vector Machine (SVM)**: Works by finding the "hyperplane that maximizes the margin between positive and negative observations for a specified class." They are convenient tools for detection of patterns, since non-separable features typically become linearly separable after mapping to a high dimension feature space [31].
- **Naïve Bayes (NB)**: Assumes that each of the features assessed is conditionally independent of one another given some class and thus, using different demographics and environmental variables as independent features makes it computationally efficient, relatively accurate, and good at supervised learning [27]. It is a probabilistic classifier, and thus outputs the category with the highest probability. However, this algorithm is sensitive to data values of zero. To resolve this issue, Laplachian smoothing can be utilized to handle categorical data, applying a small-correction to each feature count.
- **K Nearest Neighbor (KNN)**: Non-parametric, lazy algorithm that applies the assumption that similar items will exist in close proximity (measured via the distance) [27].
- **Logistic Regression (LR)**: Applies a logistic function to model a dependent variable which is binary in nature, although, it is possible to extend with alternative options as well [24].
- **eXtreme Gradient Boosting (XGBoost)**: Utilizes gradient boosted decision-trees working as an ensemble. However, instead of training the models separately it trains iteratively, working successively, so that newer models are trained as corrected or "fixed" versions of the prior ones [32].

These algorithms were chosen since they seemed to be the best fit (in terms applicable supervised machine learning algorithms balanced with computational efficiency) for our particular classification goal, using the variables from MIDFIELD. We tested different parameters for each algorithm, to find the optimal values for our data. However, we obtained the best accuracy and F1 scores using the parameters indicated below (for the algorithms where it made sense to modify them), to generate a predictive model for persistence/graduation:

- **DT**: Minsplit is the minimum number of observations that must exist in a node in order for a split to be attempted, and we tested 20, 500, and 1000. Minbucket refers to the minimum number of observations in any terminal node, and we tested 2, 100, and 500. Maxdepth is the limitation on the constructed tree's depth, and we tested 5, 6, 10, and 30. The best F1 score and accuracy was obtained using a minsplit of 20, a minbucket of 2, and a maxdepth of 30.
- **RF**: Random search (which tries random values within a range), proved more effective than a grid search. Although RF does not overfit, and the testing performance does not decrease (due to overfitting) as the number of trees increases we did observe that using 500 trees was optimal (testing ntree=50, 100, 500, and 1000).

Table 3: Comparison of Algorithms' Performance

Algorithm	Accuracy	Precision	Recall	F1 Score
DT	.8678	.8850	.9219	.9031
NB	.8225	.8368	.9120	.8728
SVM	.8527	.8528	.9420	.8952
Xgboost	.7458	.7245	**.9996**	.8401
KNN	.7538	.7587	.9259	.8340
LR	.8318	.8415	.9219	.8799
RF	**.8827**	**.8870**	.9448	**.9150**

- **SVM**: We applied SVM using Linear, Sigmoid, and Polynomial different kernels, and C values of 0.01, 0.1, 1, 10, and 20. The best performance was obtained using a polynomial kernel with a C of 20.
- **NB**: We obtained the same performance results with and without Laplacian smoothing. Therefore, applying the correction to every probability estimate did not impact the success of the algorithm.
- **KNN**: We tested K values of 1, 5, 10, 15, and 20 and obtained the best performance with 15.

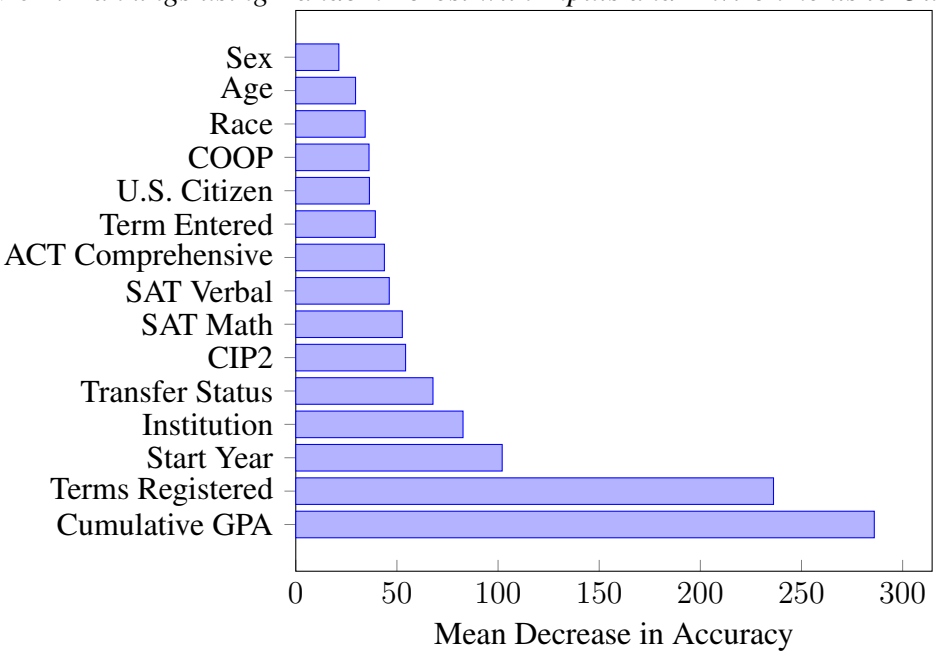
Results

Using our reduced MIDFIELD for computing students, the sample included 22% females and 78% males. Racial/ethnic group affiliation of the students were: 72.08% White, 12.19% Black, 3.81% were Hispanic/Latinx, 0.50% Native American, and 2.49% reported as "Other." When considering graduation rates for each ethnicity, we observed that the percentage of graduation from computing fields was highest amongst Hispanic/Latinx students at 50.92%, followed by Asian students at 50.20%, Other/Unknown students at 48.49%, White students at 47.48%, Native American students at 37.37%, and then Black students at 33.17%. While we did not look into why students did not complete their degrees in this particular study, it should be noted that this is consistent with the work of others, which have previously demonstrated that computing has one of the highest attrition rates, relative to other STEM fields [2].

First, we sought to determine of which ML algorithm best predicts student outcomes. Using the optimal run parameters for each algorithm, we compared the accuracy and F1 results of each, depicted in Table 3. Among the algorithms, we observed the best performance was obtained using RF, with 88.27% accuracy and the highest F1 Score, at 91.50%. Accordingly, we selected RF, to compare the variables, so that we could elucidate which were the most important in determining graduation rates.

Next, we aimed to understand what indicators predict computing students' graduation rates. To answer this we wanted to know if there is a difference, in terms of importance of the rankings, if the inputs and environment variables are run together, as opposed to when they are run separately. So, using RF, we ran three different ways, using the inputs only, using the environment only, and using both the input and environment together to predict graduation rates. RF provides the importance of each random variable, and ranks them, using a measure called the mean decrease in accuracy. *Mean decrease in accuracy* is the result of the "out of the bag" error calculation that

Figure 2: *Rankings using Random Forest with Inputs and Environments to Output*



occurs, and the greater the accuracy of the RF decreases due to excluding a particular variable, the higher the importance of that variable. Accordingly, larger mean decreases in accuracy suggest that in the data classification, a variable is more important [28, 33].

There were differences in the rankings depending on whether all variables were considered together, or separately, as determined by their mean decrease in accuracy [34]. The results from the combined RF (*Inputs* + *Environment*) $\rightarrow$ *Outputs* are presented in Figure 2, and illustrate that of all the variables, cumulative GPA was the most important, followed by terms registered and start year. Additionally, sex had the least impact on graduation rates.

Once the variables were treated separately, using a subset of the predictor variables to yield the categorical graduation rate outcome, the mean decrease in accuracy changed. The random forest rankings for *Inputs* $\rightarrow$ *Outputs* is illustrated in Figure 3, and reveals that SAT math score has the highest importance, followed by ACT comprehensive score, and then the SAT verbal score. This ordering is different from that observed when considering both input and environment together. In addition, the rankings for *Environment* $\rightarrow$ *Outputs* is illustrated in Figure 4, and although cumulative GPA still is considered the most important overall, followed by the number of terms registered, start year and institution swap in mean decrease in accuracy, and thus importance.

Discussion and Conclusions

RF outperformed other algorithms, including LR, which is the traditional analytical approach in the field of education [35–37]. As such, it was the algorithm we applied to identify which variables are the best predictors of graduation from computing. Using RF, we considered three different ways of separating the data, and the variables, to identify the most important factors. Not only did each run lead to different mean decrease in accuracy values, but also, the rankings

Figure 3: *Rankings using Random Forest with Inputs Directly to Output*

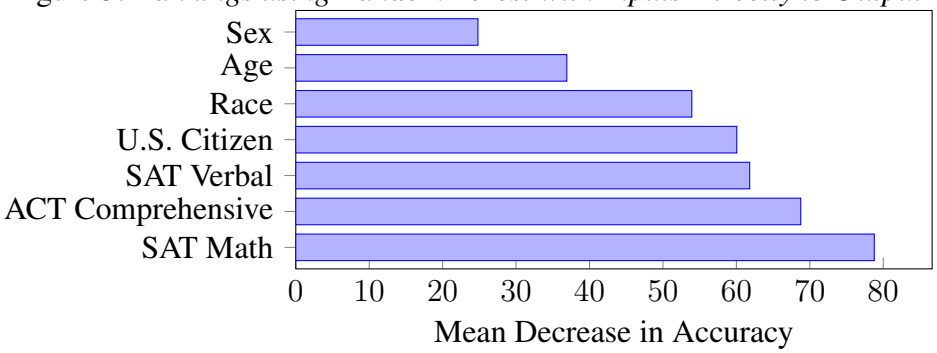
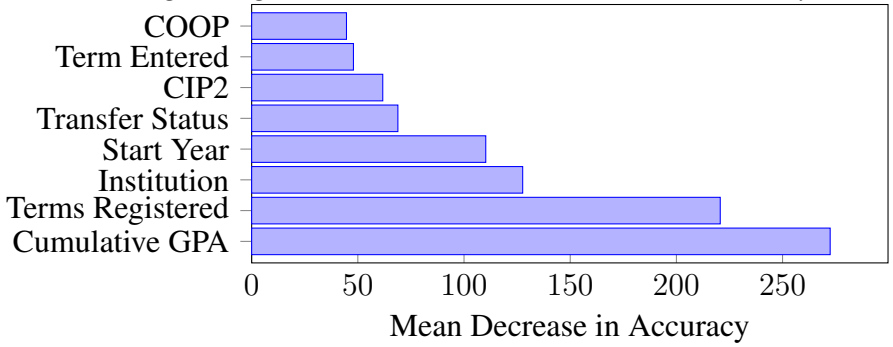


Figure 4: *Rankings using Random Forest with Environment Directly to Output*



themselves were dependent on specific combinations of variables resulting in varying importance.

Although it has been well proven that typically academic institutions offer students course performance feedback through grades [38], it is critical to understanding the role that grades may play in student outcomes. When considering (*Inputs* + *Environment*) $\rightarrow$ *Outputs*, cumulative GPA has the highest mean decrease in accuracy, at 285.98, for students' graduation in computing fields. Likewise, cumulative GPA still remains the most important variable in *Environment* $\rightarrow$ *Outputs*, although the mean decrease in accuracy here is 272.41. Since it appears that this environment variable plays such a critical role, we further analyzed the data and observed that students' above and below a GPA of 2.1 appears to be a major determinant. Considering that many schools impose a minimum GPA requirement in order to obtain their degree, this observation is in concordance with the "C" average necessitated, and makes sense why this is so crucial in its impact of graduation.

After cumulative GPA, for both (*Inputs* + *Environment*) $\rightarrow$ *Outputs* and *Environment* $\rightarrow$ *Outputs*, the variable that had the next highest mean decrease in accuracy was the number of terms registered. Results show that students who register less than five semesters (almost halfway into their career), are more likely to withdraw from the program, even though they had the required GPA to graduate. This corresponds with a high dropout rate for students enrolled in computing programs, during their initial two years. As such, it suggests that to reduce attrition, we need to better understand why students doing well enough academically do not choose to continue with

their studies in computing. Future work should look into this early time frame to perhaps include a qualitative assessment of the experiences that contribute to this lack of persistence.

Examining the rankings of *Inputs* $\rightarrow$ *Outputs*, SAT math score is considered the the primary determinant. The mean decrease in accuracy at 78.78 for the SAT math score, was the highest of the variables. This appears consistent with work which emphasizes the importance of a strong foundation in mathematics for computing interest and success [39, 40]. Interestingly, once separated out, the rankings for comprehensive ACT score rose in importance over the SAT verbal score. Unlike the SAT, which solely considers Math and Verbal, the ACT test includes four different academic skill areas: English, mathematics, reading, and science reasoning [22]. This lends itself to the explanation that for computer science students, science reasoning skills are also beneficial for academic outcomes. Considering that programming and computing heavily requires problem solving through application of specific data types, operations, and functions, so that students can perform calculations and evaluations to write, test, and then debug their code [39]- it makes sense that having a strong foundation in reasoning could be beneficial. However, we should caution that although we present what factors the analysis revealed to be most salient, we cannot infer direct causality without additional study.

This work raises several questions that future research should consider. For example, are the rankings of importance different based on personal identification with a particular gender, race or ethnicity? Likely self-identification with particular groups lends itself to unique importance of different factors. Moreover, although here we present a new methodology to the field (i.e., machine learning), how does this compare to other techniques? It might be beneficial going forward to also compare the algorithms used to more conventional methods on the same dataset. Alternatively, we could compare these same machine learning techniques on different datasets to obtain a measure of what works best overall.

Apart from the potential future directions discussed, there are several other factor-specific additions that would benefit the MIDFIELD dataset. While MIDFIELD does not include information about the availability and use of study and support resources (i.e. tutoring, mentoring, etc.), collecting information on these factors could provide interesting insight. In addition, information about what advanced placement classes were taken in a subject, or other more formative events could provide useful data to understand students' prior experiences. It should also be noted that a limitation of this work is that ideally we would want to convert all ACT/SAT scores to use the same scale. However, MIDFIELD does not report the year these standardized tests were completed. Accordingly, since the scale changed over different years, we are unable to infer exactly which each student used when completing their exam. Going forward, it would be valuable to collect this information and also to adjust the method for imputing these.

These research findings have important implications for computing students, and in understanding what qualities and characteristics before and during students' academic careers are the most important. Based on our work, cumulative GPA is critical, and a student's SAT math score and comprehensive ACT score may also play a pivotal role in predicting students' graduation from a computing field. Therefore, considering these rankings could prove beneficial to academic administrators, faculty, and other key stakeholders, to help inform and guide discussions regarding where to focus efforts to help students achieve academic success. In addition, seeing that not all ethnic/racial groups are equally represented in computing graduation, it draws attention to the

necessity of working to equalize representation in computing programs for all students.

Furthermore, employing machine learning techniques represents a novel means of exploring computing education data. Going forward, the application of supervised and unsupervised algorithms can assist researchers in uncovering novel relationships and in creating adaptive and scalable models. Whether simplifying the task of identifying students with leadership experience from a larger set of University application documents, or predicting the growth of programs based on existing information, the possibilities with ML are endless.

References

[1] U. B. of Labor Statistics, "Bureau of labor statistics, u.s. department of labor, occupational outlook handbook," Sep 2019. [Online]. Available: https://www.bls.gov/ooh/

[2] X. Chen, "Stem attrition: College students' paths into and out of stem fields. statistical analysis report. nces 2014-001." *National Center for Education Statistics*, 2013.

[3] L. Zahedi, H. Ebrahiminejad, M. S. Ross, M. W. Ohland, and S. J. Lunn, "Multi-institution study of student demographics and stickiness of computing majors in the usa," *Collaborative Network for Engineering and Computing Diversity (CoNECD)*, 2020.

[4] D. Michie, D. J. Spiegelhalter, C. Taylor *et al.*, "Machine learning," *Neural and Statistical Classification*, vol. 13, 1994.

[5] S. Hutt, M. Gardener, D. Kamentz, A. L. Duckworth, and S. K. D'Mello, "Prospectively predicting 4-year college graduation from student applications," in *Proceedings of the 8th international conference on learning analytics and knowledge*, 2018, pp. 280–289.

[6] N. S. Board, "Science and engineering indicators," *Arlington VA*, 2014.

[7] J. P. Grayson, "Race and first year retention on a canadian campus.[working paper]. york univ., toronto (ontario). inst. for social research." *Institute for Social Research*, p. 44p, 1995.

[8] L. Horn, "Placing college graduation rates in context: How 4-year college graduation rates vary with selectivity and the size of low-income enrollment. postsecondary education descriptive analysis report. nces 2007-161." *National Center for Education Statistics*, 2006.

[9] B. Brinkman and A. Diekman, "Applying the communal goal congruity perspective to enhance diversity and inclusion in undergraduate computing degrees," in *Proceedings of the 47th ACM technical symposium on computing science education*, 2016, pp. 102–107.

[10] L. V. Morris, S.-S. Wu, and C. L. Finnegan, "Predicting retention in online general education courses," *The American Journal of Distance Education*, vol. 19, no. 1, pp. 23–36, 2005.

[11] R. Boone, S. Al-Haddad, and E. Campbell, "Forecasting universities' graduation rates using multiple linear regression," in *IIE Annual Conference. Proceedings*. Institute of Industrial and Systems Engineers (IISE), 2017, pp. 902–907.

[12] O. Moscoso-Zea, P. Saa, and S. Luján-Mora, "Evaluation of algorithms to predict graduation rate in higher education institutions by applying educational data mining," *Australasian Journal of Engineering Education*, vol. 24, no. 1, pp. 4–13, 2019.

[13] H. Stumpf and J. C. Stanley, "Group data on high school grade point averages and scores on academic aptitude

tests as predictors of institutional graduation rates," *Educational and Psychological Measurement*, vol. 62, no. 6, pp. 1042–1052, 2002.

[14] G. Waugh, T. Micceri, and P. Takalkar, "Using ethnicity, sat/act scores, and high school gpa to predict retention and graduation rates." 1994.

[15] B. M. Galla, E. P. Shulman, B. D. Plummer, M. Gardner, S. J. Hutt, J. P. Goyer, S. K. D'Mello, A. S. Finn, and A. L. Duckworth, "Why high school grades are better predictors of on-time college graduation than are admissions test scores: The roles of self-regulation and cognitive ability," *American Educational Research Journal*, vol. 56, no. 6, pp. 2077–2115, 2019.

[16] G. Zhang, T. J. Anderson, M. W. Ohland, and B. R. Thorndyke, "Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study," *Journal of Engineering education*, vol. 93, no. 4, pp. 313–320, 2004.

[17] P. L. Ackerman, R. Kanfer, and C. Calderwood, "High school advanced placement and student performance in college: Stem majors, non-stem majors, and gender differences." *Teachers College Record*, 2013.

[18] A. Redmond-Sanogo, J. Angle, and E. Davis, "Kinks in the stem pipeline: Tracking stem graduation rates using science and mathematics performance," *School Science and Mathematics*, vol. 116, no. 7, pp. 378–388, 2016.

[19] A. W. Astin, "The methodology of research on college impact, part one," *Sociology of education*, pp. 223–254, 1970.

[20] A. W. Astin *et al.*, *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. Rowman & Littlefield Publishers, 2012.

[21] M. W. Ohland and R. A. Long, "The multiple-institution database for investigating engineering longitudinal development: An experiential case study of data sharing and reuse." *Advances in Engineering Education*, vol. 5, no. 2, p. n2, 2016.

[22] "Sat vs act: Which test is right for you?" [Online]. Available: https://www.princetonreview.com/college/sat-act

[23] "The integrated postsecondary education data system (ipeds) glossary. "graduation rates (gr)" definition," 2007. [Online]. Available: http://nces.ed.gov/ipeds/glossary/index.asp?id=812

[24] P. Harrington, *Machine learning in action*. 3 Lewis Street Greenwich, CT, United States: Manning Publications Co., April 2012.

[25] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.

[26] R. Kohavi and J. R. Quinlan, "Data mining tasks and methods: Classification: decision-tree discovery," in *Handbook of data mining and knowledge discovery*. Oxford University Press, Inc., 2002, pp. 267–276.

[27] S. D. Jadhav and H. Channe, "Comparative study of k-nn, naive bayes and decision tree classification techniques," *International Journal of Science and Research (IJSR)*, vol. 5, no. 1, pp. 1842–1845, 2016.

[28] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[29] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[30] G. Louppe, "Understanding random forests: From theory to practice," *arXiv preprint arXiv:1407.7502*, 2014.

[31] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. 32 Avenue of the Americas, New York, NY 10013-2473, USA: Cambridge University Press, May 2014.

[32] T. Chen and C. Guestrin, "Xgboost: reliable large-scale tree boosting system," in *Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA*, 2015, pp. 13–17.

[33] C. Zhang and Y. Ma, *Ensemble machine learning: methods and applications*. Springer, 2012.

[34] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Advances in neural information processing systems*, 2013, pp. 431–439.

[35] M. Walpole, "Socioeconomic status and college: How ses affects college experiences and outcomes," *The review of higher education*, vol. 27, no. 1, pp. 45–73, 2003.

[36] J. F. Zaff, K. A. Moore, A. R. Papillo, and S. Williams, "Implications of extracurricular activity participation during adolescence on positive outcomes," *Journal of Adolescent Research*, vol. 18, no. 6, pp. 599–630, 2003.

[37] J. L. Stephan, E. Davis, J. Lindsay, and S. Miller, "Who will succeed and who will struggle? predicting early college success with indiana's student information system. rel 2015-078." *Regional Educational Laboratory Midwest*, 2015.

[38] J. B. Main, K. J. Mumford, and M. W. Ohland, "Examining the influence of engineering students' course grades on major choice and major switching behavior," *International Journal of Engineering Education*, vol. 31, no. 6, pp. 1468–1475, 2015.

[39] T. Beaubouef, "Why computer science students need math," *ACM SIGCSE Bulletin*, vol. 34, no. 4, pp. 57–59, 2002.

[40] N. R. Zarrett and O. Malanchuk, "Who's computing? gender and race differences in young adults' decisions to pursue an information technology career," *New directions for child and adolescent development*, vol. 2005, no. 110, pp. 65–84, 2005.