Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025
Paper ID #45705

From Barriers to Bridges: A Case Study on Engineering Education

Colten Van Voorhis, Texas A&M University - Commerce

Colten Van Voorhis is a Master's student at East Texas A&M University and a Graduate Research Assistant in the Computer Science Department. My research interests include cybersecurity, autonomous driving, user privacy, and STEM diversity equity efforts.

Dr. Yuehua Wang, Texas A&M University - Commerce

Dr. Yuehua Wang is an Associate Professor of CS. Her research interests are software programming and engineering, autonomous driving, artificial intelligence, vehicular sensing and control, vehicle communication systems, resource virtualization, cyber-physical systems, Internet of Things, ubiquitous and mobile computing, wireless and mobile networks, and distributed networks and systems. She has more than ten years of research experience in academia with over twenty research publications and open-source software packages with rich programming and project development experience. She has mentored and worked closely with minority undergraduate students. She has also continuously devoted herself to advertising engineering and computer science programs and reaching out to historically underserved populations.

From Barriers to Bridges: A Case Study on Engineering Education

Abstract

Despite decades of efforts to promote diversity, equity, and inclusion in STEM education, women, minorities, black, indigenous, and people of Color students remain underrepresented in undergraduate engineering programs and STEM education even as the distribution of student demographics evolves. To examine these disparities and contribute to regional educational improvement, Texas was chosen as a model state for this case study, given its diverse and evolving K-12 student population. In Texas, between 2013 and 2022, the percentage of White students in K-12 decreased from 30.66% to 26.74%, while the Hispanic/Latino student population increased from 50.17% to 52.36%. The proportion of Black or African American students remained the same (12.50% to 12.65%) while the two or more races students increased from 1.78% to 2.89%, and Asian students changed from 3.65% to 4.68%. Compared to national enrollment by year, we found that disparities persist in Texas, where Black or African American, and Asian students are enrolled at lower percentages. In particular, Texas had a smaller population of Black or African American female students than the US (7.57% in 2013 and 7.37% in 2022), with only 6.61% in 2013 and 6.20% in 2022 in Texas. The underrepresentation raises concerns about barriers deterring students from minority groups from pursuing engineering degrees, a challenge observed across multiple states despite varying demographic compositions. In this paper, we collect data and conduct a detailed case study by examining university data on engineering education, including enrollment, retention, and degree completion. We are dedicated to identifying the primary factors and barriers that influence the education of minority students in the field of engineering. Machine learning is introduced and used as a key tool to analyze the collected data and predict future trends in engineering education. Results, challenges, and future efforts are then discussed. The findings from this study provide insights into existing challenges and propose strategies to address barriers and promote a more inclusive engineering education landscape.

Introduction

The past few decades have witnessed an unpredictable and rapidly changing world fueled by high technologies. As the world's leading global leader in science and technology, the United States has invested the most in research and development and awarded the most advanced degrees [1]. Although absolute science and technology levels continue to increase, the relative share of global science and technology activities in the US is seen to be declining, according to the 2024 State of U.S. Science and Engineering issued by the National Science Board [2], [3]. This is mainly due to the shortage of skilled technical workforce in the fields, who have emerged as the driving force for remarkable achievement and innovation in the United States. A study conducted by the Manufacturing Institute and Deloitte [4], [5] predicts that there might be 3.8 million unfilled positions across the entire U.S. between 2024 and 2033, and approximately half of these available positions (1.9 million) will remain unfilled if the skills and applicant gaps cannot be filled. There is an urgent necessity to enhance university enrollments and degree completion to help fill the gaps and meet the demands of high-skilled professionals.

This shortage of skilled technical workforce with university degrees can be attributed to the lack of interest in higher education and careers in STEM (Science, Technology, Engineering, and Mathematics)-related fields at both K-12 and college levels. The latest State of U.S. Science and Engineering [6] and NSF BPCnet's public data [7], [8] show that women, minorities, and BIPOC (Black, Indigenous, and People of Color) people are underrepresented in K-12 education and among bachelor's degree holders in science and engineering fields. In 2022, the percentages of K-12 students who identify as Black or African American (15.09%), Hispanic/Latino (26.89%), American Indian or Alaska Native (0.98%), and Asian (5.12%) remain below the national average of 47.79% for White students. States overall face varying circumstances in K-12 and STEM education. Compared to the national enrollment by year, we identified lower percentages of Asian and Black or African American students in Texas. Specifically, the percentage of Black or African American female students in Texas was 6.61% in 2013 and 6.20% in 2022, which is far lower than the national average of 7.57% in 2013 and 7.37% in 2022. Other demographics, except Hispanic/Latino, experience similar situations in Texas.

On top of that, pursuing higher education in STEM fields after graduating from high school to college is challenging, as students face increased academic demands, greater independence, and the pressure towards earning a degree. Only 333,000 (18%) out of 1.8 million bachelor's degrees were awarded in STEM in the cohort year 2015-2016, although there was 3.1 million youth aged 16 to 24 who were high school graduates [9] and the national high school graduation rate hit 71.7% in 2011-2012. In fall 2012, 7.2 million (40%) undergraduate students [10] were enrolled in 2-year institutions (also called community colleges) for certificates and associate degrees, and 10.6 million (60%) were enrolled in 4-year colleges and universities for bachelor's degrees. However, within 6 years, only 5% of undergraduate students who enrolled in community colleges had received a certificate, and 22% had received an associate's degree, while there were 51% of undergraduate students who enrolled in 4-year colleges and universities and received a bachelor's degree. It has been clearly observed that the persistence, retention, and attainment of STEM undergraduate students in community colleges and 4-year universities remain low while gender disparities [11] in STEM education persist.

To promote equity, diversity, and inclusion in STEM education, in this study, we independently collect university data from the state of Texas to investigate the intersection of demographic trends, institutional funding, and program effectiveness and the effect these factors have on STEM program growth using machine learning. The machine learning algorithms used are decision tree, adaptive boosting, random forests, residual neural network, long short-term memory, convolutional neural network (CNN), transformer, and a standard linear regression model. To avoid overfitting, we first group the universities into subcategories based on the sum of squared errors (SSE). The universities in the same group share various similarities, including comparable funding levels, institutional policies, and program structures geared at assisting underrepresented groups in STEM disciplines. Based on the clustering, we report our experimental results and research findings performed using various metrics, including R^2 score, MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and MAE (Mean Absolute Error). The results can assist in emerging quantitative

Table 1: Efforts to Improve STEM Enrollment

Activity	Implementation
STEM Learning Community	Women in STEM, SLC program, first-year interest groups
Living Communities	On-campus housing specific to STEM majors
STEM Specific Honors Program	STEM based honors courses and thesis options
Career Preparation and Networking Opportunities	Career centers, internship pipelines, industry partnerships
Peer Mentorship Efforts	Mentorship programs, classroom visits, campus career panels
STEM Research Opportunities	Freshman research initiatives, summer undergraduate research programs, department research
Faculty Professional Development	Workshop and consultation opportunities, career advancement resources, development grants
Targeted STEM Orientation	Department Orientation Sessions and freshman programs

research that centers on enhancing STEM programs and improving STEM participation and completion for women, minorities, and BIPOC. To our knowledge, this is one of the few quantitative studies contributing to identifying primary program factors, understanding their impacts, and introducing machine learning methodologies as a tool to begin addressing barriers to STEM education enrollment and retention.

The organization of this paper is as follows. In Section II, we provide an overview of positive institutional activities and their implementation to promote minority undergraduate education. Based on that, we collect data and conduct data analysis and processing in Sections III and IV. The machine learning algorithms are studied, explained, and evaluated using our dataset in Sections V and VI. We finally conclude our presentation in Section VII with future directions.

Background

To address barriers present in undergraduate STEM education, existing positive institutional activities and their implementations have been summarized, including STEM Learning and Living Communities, STEM Honors Programs, Career Preparation and Networking Opportunities, Peer Mentorship Efforts, STEM Research Opportunities, Faculty Professional Development, and Targeted STEM Orientation listed in Table 1.

Research Program and Opportunities Perhaps the most widely adopted, though indirect, method of increasing STEM enrollment would be an institution's ability to offer research opportunities in STEM fields. Research work allows students to connect with campus faculty and, more often than not, a group of like-minded peers to work towards a common goal. Depending on a university's status and local relationships, companies directly involved in science/technology industries might reach out to sponsor or contribute to these research opportunities, giving students the potential to network with different companies and feel more connected to their academic environment while working towards the completion of some goal. Although student involvement in research has been shown to typically lead to higher retention and graduation rates [12][13], it isn't uncommon for smaller institutions to lack the resources necessary to offer meaningful research programs for their student population. A campus working towards offering more research opportunities, regardless of size, would be working towards elevated STEM numbers by extension.

Faculty Development While research opportunities offer a relationship between faculty and students that typically benefits the student the most, development opportunities that benefit university staff members are another worthwhile activity that have been shown to lead to an increase in STEM numbers. Universities that offer staff the chance to participate in specialized programs, workshops, and mentorships typically see their efforts reflected in higher levels of student engagement and a greater willingness to participate in community efforts [14][15]. Members of the National Center for Faculty Development and Diversity (NCFDD) produce staff that utilize highly effective evidence-based teaching methods, are more likely to emphasize inclusivity and equity within their classes, are better equipped to interact with and advise students, and are overall more confident in their ability to teach, all of which contribute positively to student retention. Faculty development initiatives allow educators to learn techniques to encourage inclusive classrooms and support various learning needs [16]. Community outreach efforts and K-12 engagement also contribute greatly to STEM interest early on and create a pipeline for motivated students to enter STEM fields. By combining local efforts with faculty growth, institutions can effectively handle some easily addressable structural and social challenges impacting student success in STEM to improve attraction, persistence, and retention rates of students.

Career Preparation and Networking They are essential and highly valuable for STEM students [17], [18]. Career centers can offer students the opportunity to connect with companies and potential employers for employment and internship opportunities. Implementing these opportunities can range from offering a career planning class for upper-level students to giving informational seminars highlighting the career opportunities available for each major. Additional efforts can even be directed to individual demographics such as women in [19]. Regardless of how they are implemented, the most appealing factor of these efforts is that STEM students involved in campus-sponsored career development programs perform higher academically than those not involved [20]. Networking programs, in particular, present the opportunity to boost retention rates at a relatively low cost easily.

Honors Programs Offering an honors program specific to STEM students would give those who participate the opportunity to immerse themselves in content that directly covers the technical aspects of STEM majors in ways that a more generalized honors program wouldn't typically. In concept, this would create a pipeline for student participation that directly leads to future research work or internship participation if such requirements are implemented. The benefits of research work and the networking that comes from internship efforts would both apply as previously discussed, and student participants would feel a much greater commitment to their cohort. Overall benefits are most prevalent for underrepresented groups and female students [21]. Studies on the topic observed that honors program participation is directly associated with first-year GPA at less competitive institutions, retention to the third and fourth years, and graduation rates at the completion of the fourth year. It follows that growing universities looking to increase minority success across all majors, not just STEM, have reason to believe that honors programs specific to the major could greatly contribute. It was also reported that fourth-year GPA averages were higher specifically for black students, first-year satisfaction rates were greater for female students, and third-year retention was exceptionally high for those students with lower parental education.

STEM Learning Community Learning communities facilitate social networking and the creation of study groups contributing to an increased sense of belonging within STEM majors by having students take common classes upon their first semester. The efforts of the Memphistep program at the University of Memphis aim to see greater retention rates reflected through enrollment numbers over time, specifically for minority and at-risk students by establishing LLCs for these groups [22].

Additional studies on the MemphiSTEP program indicate that those involved with the program's activities, more specifically those involved with academic preparation and community involvement, displayed noticeably elevated retention and performance rates when directly compared with peers outside of the program [23]. Outcomes of math bootcamps indicate that the average GPA rose from 2.53 to 2.73 while retention rates rose from 57% to 80%. Outcomes of the various networking efforts increase the average GPA from 2.53 to 2.76, with retention going from 57% to 2.76%. Data on year-over-year demographics indicate that at-risk groups benefited the most from the MemphiSTEP program, with Black Freshmen, in particular, seeing an average GPA increase from 1.85 to 2.43, with retention going from 44% to over 81%. When considering factors such as gender, race, academic standing, and prior performance, this outcome supports the idea that reteretention-oriented programs can have a direct, positive impact on diversity within programs and overall completion rates intention oriented programs can have a direct, positive impact on diversity within programs and overall completion rates of STEM fields. As a whole, MemphiSTEP and similar LLC efforts showcase how effective the implementation of thorough support systems can be at improving the experience of STEM students and alleviating some of the most commonly encountered academic and social challenges. Such implementations validate the idea that STEM retention rates can ultimately be enhanced at the institutional level depending on what programs a campus is willing to implement.

Targeted STEM Orientation STEM-based orientation programs can introduce students to the demands and rewards of STEM fields early on, while effective intervention systems can help faculty advisors identify students who might be struggling [24]. Mentorship programs work to combine exposure to STEM concepts and implementation of a safety net, benefiting minority groups within STEM, those who typically have fewer role models to look to. Targeted orientation programs offer incoming freshman and transfer students the opportunity to engage with both the structure and material of STEM courses prior to full-time enrollment [25]. At face value, this gives students a glimpse at some of the content they'll be studying once prerequisite requirements have been completed. However, programs that extend throughout the initial year of enrollment have the time to identify those students who would be considered at risk of dropping out of STEM programs and implement remediation efforts to prevent the switching of majors [26]. Early adoption of students within STEM programs has the added benefit of allowing students to interact with future professors and maximize their interactions with cohort members.

Prior studies on improving underrepresented and minority student graduation rates found that students who aspire to pursue a graduate-level degree eventually see increases in undergraduate completion rates by over 30%. Exposing minority students to the potential benefits of university graduate programs early on can increase the number of students seeking graduate degrees, thereby increasing retention [27]. Furthermore, underrepresented students who joined a pre-professional or departmental club were found to increase their rate of graduation by over 150%. A university's ability to attract incoming students to STEM clubs and similar programs stands a much better chance of increasing STEM numbers.

Methods

This study examines the enrollment and completion rate of minority groups in Texas as a representative/model state, with the potential for extension to other geographic regions. We first collect university data and perform exploratory analysis and processing to reveal insights within the data. Finally, machine learning algorithms are studied to build regression models whose efficacy can be evaluated and compared to best predict future trends in STEM programs.

Dataset The growing population of minorities attending Texas institutes reflects changes to the statewide population as a whole and heavily contributes to documenting future trends. Data was collected and combined based on public, biyearly enrollment, and degree award statistics from 2012 to 2022 from a range of database sources, including the National Science Foundation By the Numbers site, the IPEDS Engineering Degrees Awarded site, and the Engineering PLUS Metrics Landscape site for grant, graduation, and enrollment information of 28 institutions respectively. A total of 168 entries from 28 institutions were included, and 41 different attributes were used for each entry.

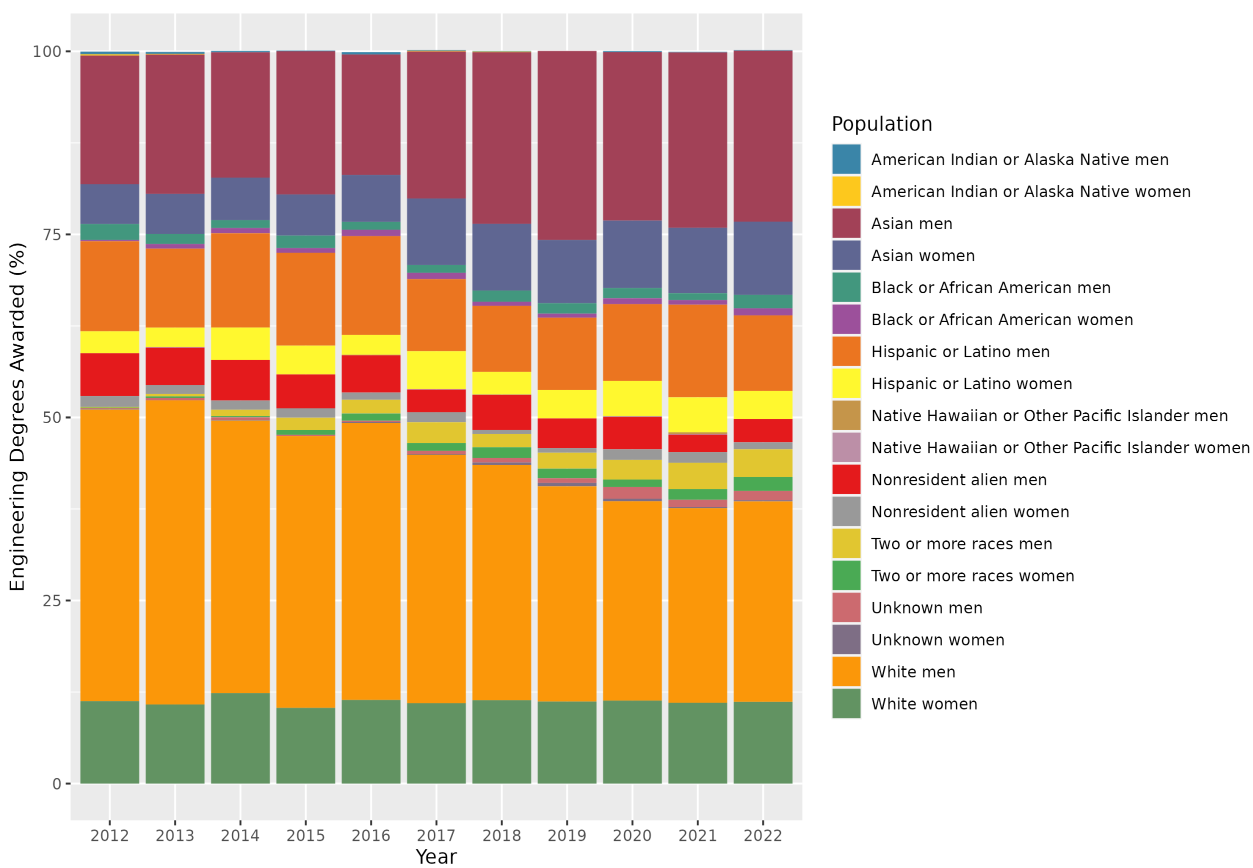

Figure 1: Example Univesity Enrollment by Ethnicity

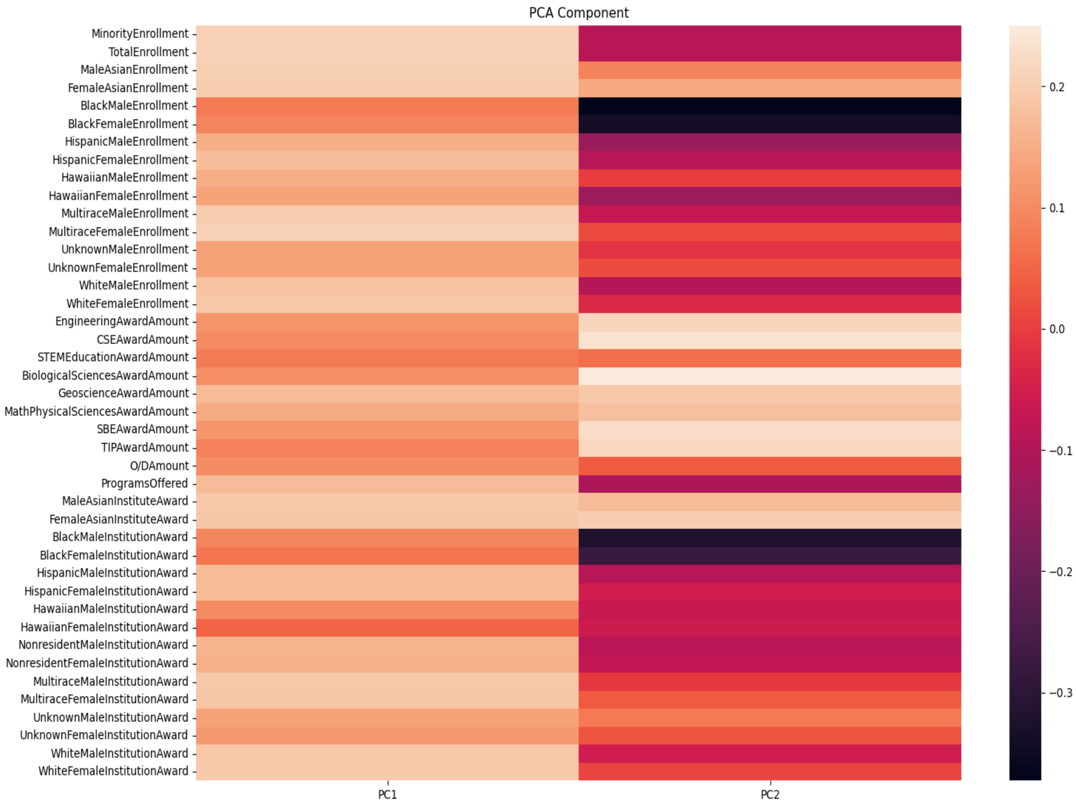

Figure 2: Data Features Mapped to PCA

Ethnicity numbers are based on male and female enrollment for Asian, African American, Hispanic, Nonresident, Multirace, Unknown, and other demographics ,as shown in Figure 1. The spreadsheet includes data collected from the institutions including Corpus Cristi, Texarkana, Kingsville, West Texas,Prairie View, Commerce, and International campuses of the Texas A&M university system, the Arlington, Dallas, San Antonio, Tyler, Rio Grande Valley, El Paso, and Permian Basin campuses of the University of Texas school system, as well as Midwestern State, Tarleton, North Texas, and Houston-Clear Lake universities. Additional institutions include Rice, Southern Methodist, Baylor, Texas Tech, Houston, Texas Christian, Texas State, Lamar, A&M College Station, and UT Austin universities.

In addition to enrollment numbers by demographic, data was compiled based on NSF funding by STEM directorate, including Engineering, Biological Sciences, Computer & Information Science & Engineering, STEM Education, Geosciences, Mathematical and Physical Sciences, Office of the Director, Social, Behavioral, & Economic Sciences, and Tech, Innovation, & Partnerships for a total of twenty-eight Texas universities. This was done to observe data trends for institutions with different NSF funding rates so that these trends could be extrapolated to additional institutions and states. Data taken from all universities within the dataset was used as input to construct prediction models for future enrollment numbers. Data itself was normalized so that the machine learning algorithms utilized could more efficiently identify trends and relationships between variables.

Raw Data Analysis and Processing Principal Component Analysis was utilized to reduce the high dimensionality of the initial dataset down to a handful of abridged dimensions for simplified visualization of the clustering process. It reduced data features to two principal components that capture the highest variance within the data itself, as shown in Figure 2. Variance is the measure of the degree of spread between a data point and its mean value. Negative values for PCA components are not correlated with negative variance; instead, they denote how one data entry relates to another in terms of all features taken together. Data was first normalized using a standard scaler so that all features were translated onto the same scale with a mean of 0 before a covariance matrix moved the data to a new coordinate system in which the axes correspond to directions based on principal components 1 and 2 [28]. PCA1 and PCA2 were selected to capture the most significant variance in the data and offer a useful visual representation for a better understanding of clustering results.

Labeling of the data was conducted using a combination of clustering methods in order to identify additional distinctions between universities within the dataset. Based on the classification of universities as research institutions in the Carnegie Classifications of Institutions of Higher Education, it follows that the Texas schools within the dataset fall into one of three categories based on the amount of funding they received and the research opportunities they're able to pursue. We continued to use three categories when factoring in graduation and enrollment rates for minority students to keep some level of consistency for the number of clusters to use. Cluster labels were assigned using DBSCAN, Agglomerative (AGC), and BIRCH clustering techniques and averaged to create finalized categories that were then used to train the previously mentioned machine language models. Additionally, feature selection was made in an effort to reduce any interdependencies that exist between features with numerical values. Out of the 41 total features present in the dataset, 25 were identified as being the most influential based on the understanding that the model coefficients of these features produced a score greater than 1. Comparing Figure 3 to Figure 4 shows how clustering scores could differ based on what features were deemed more influential than others.

Universities with cluster label zero (C0) were identified to have cluster scores with absolute values of components being slightly greater than 0, indicating a group of universities that have increased enrollment numbers, minority enrollments, or award distributions, as shown in Figure 3. The features for several institutional awards related to larger minority groups and some of the more technical funding categories tend to reflect positive values, suggesting these universities provide moderate support in these areas. Examples of this cluster include schools ranging from UT Dallas or Tarleton to institutes such as Texas State University or Univerthe sity of Houston. Universities that fall into this category are most commonly associated with both R1 and R2 research institutions but are limited by the total amount of funding they receive or by relatively lower levels or overall student enrollment. Alternatively, schools with cluster label one (C1) were identified to have cluster scores whose absolute value tend to be closer to 0, suggesting universities with lower to moderate total enrollment, minority enrollment, or funding. The cluster can be characterized by average values for most of the awards and student body in general. Examples of this cluster include Midwestern State University and A&M Texarkana. A mix of some R1 and non-research schools would be included here due to low enrollment or grant totals when compared to cluster zero schools. The final cluster label two (C2) was saved for schools whose absolute value component scores are observably higher than 0 and include all entries associated with A&M College Station and the University of Texas at Austin. These high scores indicate universities with larger

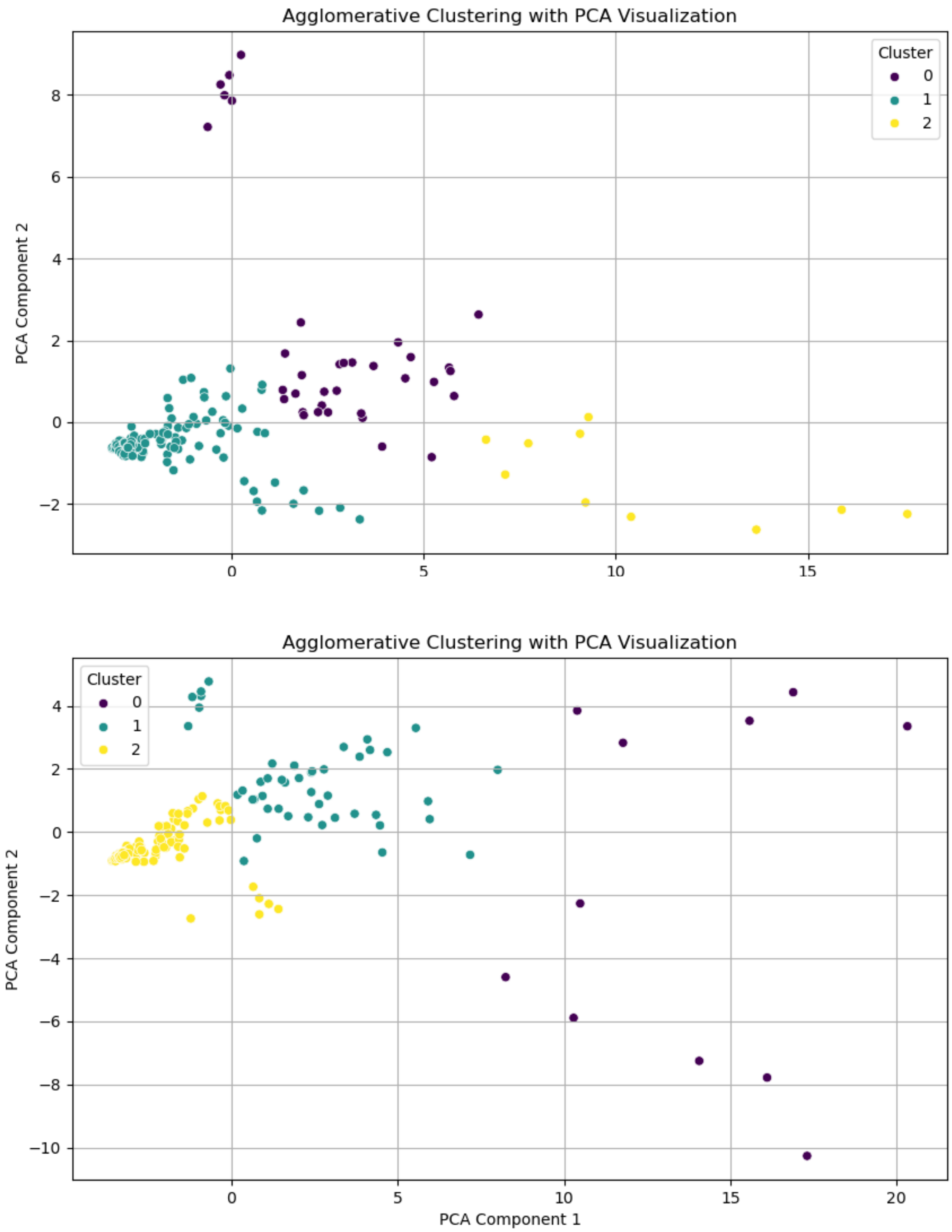

Figure 4: AGC without Feature Selection

total enrollment and minority enrollment, and provide more significant award amounts and support, particularly in funding categories that see decreased award amounts in comparison to other clusters. This group [29]represents universities with more substantial funding and a more significant proportion of both enrolled and graduating minority students.

<u>*Machine Learning Algorithms*</u> Eight(8) machine learning algorithms were studied and utilized

to create regression models for prediction based on the training and testing data. The models are Linear regression [30], Convolutional Neural Network (CNN) [31], Adaptive Boosting (AdaBoost) [32], Decision Tree (DT) [33], Random Forest Regressor (RFR) [34], Residual Neural Network (ResNet) [35], Long Short-Term Memory (LSTM) [36], and Transformer [37].

Linear regression is a statistical approach for modeling the relationship between a dependent variable and one or more independent variables. It involves fitting a straight line that minimizes the difference between observed and predicted values. The CNN architecture includes convolutional layers, pooling layers, and fully connected layers. Convolutional layers use kernels to extract hierarchical features by performing element-wise multiplications over spatial dimensions. CNNs are good at capturing spatial features, making them powerful for tasks involving spatial dependencies. A Decision Tree is a non-parametric model used within regression tasks by recursively splitting data into subsets based on feature divides. The overall simplicity of the model typically makes them effective for documenting non-linear relationships but can fall short due to overfitting. AdaBoost combines multiple weak learners, like decision trees, to create a robust predictor. The final model aggregates predictions as a weighted sum of weak learners, improving accuracy iteratively. AdaBoost's loss function, often weighted MSE, penalizes larger residuals from harder examples more heavily. ResNet is a deep learning model created with the intention of addressing the vanishing gradient problem common within deep architectures. ResNet comes equipped with skip connections meant to bypass layers by adding the input directly to the output of those layers facilitating for gradient flow during backpropagation. LSTMs are specialized Recurrent Neural Networks designed to handle long-term dependencies in sequential data. Each LSTM cell includes input, forget, and output gates that regulate information flow, enabling memory retention over extended sequences. Sequential inputs are processed with cell states and hidden states that update at each time step. LSTMs are particularly suited for time-series regression due to their ability to capture temporal patterns.

Transformers revolutionize sequence modeling by employing a self-attention mechanism to capture dependencies between elements in a sequence, irrespective of distance. Input elements are transformed into query, key, and value vectors, with self-attention calculating the relevance of each element via scaled dot-product attention. Layers stack attention mechanisms and feedforward layers, often with ReLU or Gaussian Error Linear Unit (GeLU) activations. For regression, the final layer uses linear activation for continuous outputs. Transformers can deal with long-range dependencies, making them ideal for sequential and structured data regression tasks, with loss functions like MSE guiding parameter optimization. It scales better with increasing data size and sequence length, while CNNs and LSTMs can struggle with very high-dimensional or long-range data. Understanding the above distinctions helps in selecting the most appropriate model for specific data types and task requirements. Table 2 shows the configuration of machine learning algorithms studied in this paper.

In Table 2, the linear regression model was employed to provide baseline results that could be compared to more complex models. CNNs and Transformers employ ReLU ($\text{ReLU}(x) = \max(0, x)$, where $ReLU(x) = x$ if $x >= 0$ and $ReLU(x) = 0$ if $x < 0$) for computational efficiency and gradient stability, with Transformers occasionally using GeLU ($\text{GELU}(x) = \frac{x}{2}\left(1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$, where $\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt$) for probabilistic modeling. LSTMs utilize sigmoid for gating mechanisms (input, forget, output) and tanh for cell state updates, ensuring

Table 2: Model Configurations

Model	Configurations
Linear Regression	Solver: Auto, Fit Intercept: True, Normalize: False, Max Iterations: 1000
Adaptive Boosting	Estimators: 50, Learning Rate: 1.0, Loss: Linear, Random State: 10
Decision Tree	Criterion: Mean Squared Error, Max Depth: None, Min Samples Split: 2, Min Samples Leaf: 1
Random Forest	Estimators: 100, criterion: Mean Squared Error, Max Depth: None, Min Samples Split: 2, Min Samples Leaf: 1
CNN	Number of Layers: 4, Kernel Size: 3x3, Strides: 1, Activation Function: ReLU, Optimizer: Adam, Batch Size: 3, Learning Rate: 0.001
LSTM	Number of layers: 2, Units per layer: 50, Activation: tanh, Optimizer: Adam, Batch Size: 32, Learning Rate: 0.001, Return Sequences: False
Transformer	Attention Heads: 8, Layers: 6, Feed and Forward Size: 2048, Dropout Rate: 0.1, Learning Rate: 0.001
ResNet	Number of Layers: 64, Activation Function: ReLU, Optimizer: Adam, Batch Size: 32, Learning Rate: 0.001

bounded state transformations. Loss functions ensure that each model optimizes predictions according to the problem's demands, whether by penalizing large deviations or focusing on specific data aspects. CNNs, Transformers, and LSTMs use Mean Squared Error (MSE) to penalize large residuals and Mean Absolute Error (MAE) for robustness against outliers. AdaBoost emphasizes weighted error minimization to focus on challenging examples. R squared score is used to measure the proportion of variance in the dependent variable that is explained by the model, and is typically indicative of how well the model describes changes in the target variable. Root Mean Squared Error (RMSE) functions similar to MSE and is used to provide results in the same units as the target variable and provides additional interpretations of model results.

Results and Discussion

The dataset was split into training and testing sets where eighty percent of the dataset was used for training models with the remaining twenty percent being used to gauge the prediction accuracy of those models. The same train/test split was used with all models in order to directly compare metrics seen in Figure 5. A validation split was not utilized for the initial dataset due to its overall small size and limited entries. Future efforts could build upon what was accomplished here by adding additional entries to the dataset by incorporating university data from other states and including a validation set of data.

Dataset features were used to train models with the goal of predicting values for the percentage of minority students enrolled at Texas institutions in addition to the ratio of minority graduations to total enrollment numbers. Training based on individual cluster groups, as well as the dataset as a whole, was conducted in order to document more nuanced comparisons.

MSE, RMSE, and R^2 (R Squared) are metrics used to evaluate the performance of ML algorithms and optimize models. R^2 can be used to evaluate how well a model captures the variability

Table 3: Minority Enrollment with Clusters in terms of MSE

Model	C0	C1	C2	Mixed
Baseline	81.927	198.814	**7.910**	229.909
Adaptive Boosting	**10.511**	25.638	16.460	50.488
Decision Tree	**6.740**	72.764	120.009	88.754
Random Forest	16.259	33.366	**4.526**	24.443
CNN	**14.379**	78.031	30.801	52.199
LSTM	0.074	0.098	**0.070**	0.085
Transformer	**0.017**	0.138	0.338	0.063
ResNet	0.078	0.087	0.422	**0.041**

Table 4: Minority Degree Completion with Clusters in term of MSE

Model	C0	C1	C2	Mixed
Baseline	0.00190	**0.00110**	7.07600	0.00150
Adaptive Boosting	0.00030	0.00023	**0.00011**	0.00094
Decision Tree	0.00230	**0.00094**	1.65400	0.00180
Random Forest	0.00060	**0.00017**	**0.00017**	0.00093
CNN	0.00050	**0.00037**	4.72200	0.00046
LSTM	0.42400	0.15900	0.14400	**0.05700**
Transformer	0.42000	0.95300	**0.01030**	0.54960
ResNet	0.50600	**0.17800**	0.20700	0.21300

of a target variable, with coefficients falling between 0 and 1 with a score closer to 1 indicating a higher accuracy model. MSE is the difference between predicted values and actual values between the training and testing sets with values closer to zero indicating more accurate predictions. RMSE functions are similar to MSE but express errors in the same units as the target variable, meaning that the error result is more interpretable. Like with MSE, values for RMSE closer to zero indicate more accurate predictions. Mean Absolute Error measures the average magnitude of error between predicted and actual values, regardless of whether the error is itself positive or negative. Values closer to zero [38] indicate a greater prediction accuracy.

Table 3 shows the average MSE scores across clusters with the ratio of minority enrollment to the total of enrollment as Target. In this case, the utilization of trained learning models typically produced more accurate predictions than the use of the baseline model alone. This holds true for Cluster Zero (C0), Cluster One (C1), and the Total Dataset groupings with learning models producing mean squared errors that were noticeably lower than baseline numbers. Deep learning models were particularly effective for predictions with C0 and C1 with the use of the Total Dataset producing less accurate results despite still being relatively high. Cluster Two (C2), however, saw inaccurate predictions based on all metrics for three of the seven models, with the remaining two models producing output that was functionally similar to baseline results. The reduced size of C2 likely contributed to difficulties in developing accurate models leading to RandomForest and LSTM seeing higher accuracy due to their architectures. Overall the Transformer was most effective for C0, the ResNet was most effective for C1 and the entire dataset, and the LSTM was

Table 5: Minority Enrollment

Models	$R^2$1	$R^2$2:	MSE1	MSE2	RMSE1	RMSE2	MAE1	MAE2
Baseline	0.578	0.247	129.640	93.683	10.282	6.679	8.027	5.045
Adaptive Boosting	0.780	0.444	25.714	17.502	4.862	2.863	4.153	2.365
Decision Tree	0.634	0.395	70.417	33.875	7.563	3.850	5.714	2.618
Random Forest	**0.903**	**0.461**	19.566	12.379	4.176	2.448	3.286	1.915
CNN	0.619	0.392	40.989	24.702	6.231	3.476	5.186	2.734
LSTM	0.857	0.451	**0.080**	0.041	**0.280**	0.143	0.354	0.160
Transformer	0.608	0.436	0.136	0.041	0.327	0.132	**0.256**	**0.094**
ResNet	0.517	0.439	0.159	**0.039**	0.351	**0.125**	0.296	0.099

Note: Notes:$R^2$1 and $R^2$2 are the total and weighted average of R^2, MSE1 and MSE2 are the total and weighted average of MSE, RMSE1 and RMSE2 are the total and weighted average of RMSE, MAE1 and MAE2: are the total and weighted average of MAE.

Table 6: Minority Degree Completeion

Models	$R^2$1	$R^2$2:	MSE1	MSE2	RMSE1	RMSE2	MAE1	MAE2
Baseline	0.347	0.136	0.001	0.001	0.031	0.019	0.024	0.015
Adaptive Boosting	**0.772**	0.366	**0.001**	**0.001**	0.019	0.012	**0.014**	0.009
Decision Tree	0.194	0.091	0.061	0.022	0.034	0.020	0.026	0.015
Random Forest	0.726	0.370	0.169	0.059	0.020	0.011	0.015	0.009
CNN	0.770	0.388	0.179	0.062	**0.019**	**0.010**	0.015	**0.008**
LSTM	0.749	**0.411**	0.237	0.092	0.418	0.182	0.349	0.147
Transformer	0.404	0.138	0.522	0.324	0.611	0.381	.0458	0.277
ResNet	0.661	0.369	0.279	0.130	0.503	0.248	0.429	0.201

most effective for C2.

In the case where the prediction target was based on the ratio between minority graduations and total enrollment, trained models seemed to be less effective in comparison to the alternative target, as shown in Table 4. While predictions continued to outperform the baseline model, model metrics seemed to vary more widely than in previous cases. Baseline models were able to produce fairly accurate predictions across all clusters save for C2. Adaptive Boosting, Decision Tree, and Random Forest alternatively were able to produce MSE scores that would be considered highly accurate in comparison to all clusters. Despite being more accurate than baseline predictions with the alternate target, the deep learning models utilized seemed to struggle in this instance with predictions ranging across cluster groups. The outliers within C2 across models justify the idea that the small number of data within the cluster benefits from the use of a linear regression model similar to the type used for baseline predictions. Methods that produced output similar to baseline seemed to excel for this reason. As a whole predictions made using the entire dataset produced better results than those made using individual clusters.

Table 5 and 6 show the performance of models in terms of R^2, MSE, RMSE, and MAE. Random Forest, LSTM, and CNN are still showing better performance compared to other models. When comparing factors that contribute to differences in STEM enrollment numbers across universities, it's worth noting the discrepancies between minority-serving institutions (MSIs) and non-MSIs that tend to directly influence student outcomes. While MSIs are responsible for serving a higher proportion of underrepresented students compared to their counterparts, they tend to receive lower funding amounts across all categories. Such a difference limits an institution's ability to provide extensive research opportunities, advanced campus facilities, and support programs crucial to fostering STEM success in regards to minority populations. Institutes that receive higher levels of funding typically see the use of more effective tools for attracting and retaining students to STEM programs through the use of tailored scholarships and mentorship efforts. While there are certainly outliers within the Texas university dataset we observed, UT Austin and A&M College Station for example, the dataset as a whole details the need for target funding policies for universities with high minority enrollment numbers.

Regional comparisons of enrollment trends and funding reflect the idea that geography and location can be impactful to STEM education and long-term success [39][40]. A state such as Texas demonstrates a range of funding distributions that do not directly align with demographic needs based on population diversity. A great example of this concept would be urban institutions tending to receive higher rates of NSF funding in comparison to rural or more isolated campuses, even when these campuses are responsible for serving a larger overall proportion of minority populations in comparison to more prominent universities [41]. Imbalances like this compound upon rural institutes and create additional challenges for schools already struggling to compete with more notable STEM programs. Efforts to address regional funding differences could improve minority access to quality STEM educational opportunities and improve enrollment rates at smaller institutions.

It is worth noting the direct impact funded grants have on long-term engineering enrollment and graduation rates, especially for those institutions that effectively utilize these resources year after year[42][43]. With prominent programs using these grants to implement additional research opportunities, create more nuanced learning communities, and fund faculty development, it follows that such universities would see higher STEM engagement and program completions. Programs receiving smaller, inconsistent funding amounts tend to see greater fluctuations in their enrollments and completions as reflected in the dataset numbers, reflecting the idea that sustained investment is an important component of improving a STEM program's metrics[44]. Changes in funding over time produce direct results on enrollment and graduation trends, with increased funding producing greater minority enrollment growth and graduation rates over time and decreased funding producing having the opposite effect. The cumulative effect of funding fluctuations means that institutions able to operate within periods of sustained funding yield more stable improvements to STEM metrics as a whole [45].

When looking at individual demographic groups, targeted interventions could be used to address barriers unique to specific groups[46]. Where a historically underrepresented group, such as African American women would stand to benefit the most from mentorship programs or support networks meant to encourage inclusion, another group, such as Hispanic students, might stand to gain more from a learning community that highlights their cultural origins. Depending on the university, focusing on the necessities of specific demographics might accomplish more for STEM growth than looking to boost metrics for all groups [47].

Where the models used within this study are concerned, optimizing hyperparameters for each

Table 7: Hyperparameter Tuning

Cases	R2 Score	MSE	RMSE	MAE	No. Filt.	Size (Filt.)	Layers
Best Case	0.889	41.690	6.457	4.948	128	5	3
Default	0.761	53.656	7.286	5.418	64	3	4
Worst Case	0.676	120.437	10.974	8.011	16	7	3

model would, in theory, produce results that are slightly more accurate when compared to the default configurations used within this paper. The primary example seen in table 7 shows how much of an impact changing the number of filters, size of filters, and number of layers can have on the resulting metric scores of the CNN model. From the limited optimization testing conducted, the R2 score and MSE were seen to improve by over 0.1 and 10 points, respectively, in the best case, as shown in Table 7. It's also worth noting how detrimental improper tuning can be to metric scores as well. As a whole, the number of filters used and the size of each filter seemed to be more influential than changing the number of layers used.

Conclusion

In order to better interpret STEM program growth over time, it's necessary for future research efforts to explore additional metrics for gauging STEM program success as well as employ alternative models for prediction analysis. Using metrics that go beyond enrollment and graduation rate, such as academic resilience, student engagement, and career trajectory post-graduation, would, in theory, paint a clearer picture of a program's effectiveness and provide additional variables for more accurate predictive models. The utilization of more advanced machine learning models or additional models could lead to predictions that are more accurate on average and provide deea per nuance of complex relationships between factors influencing STEM outcomes. The use of additional universities from states outside of Texas, the expansion of the dataset to include years beyond the range used within this dataset, and the inclusion of additional grant statistics could all lead to the development of more accurate models as well. The use of anecdotal evidence could also be utilized alongside quantitative metrics to more clearly define the state of an institution.

Cultural shifts within STEM departments have the potential to positively address disparities within engineering education. STEM programs as a whole must transition away from placing high regard on academic difficulty as a barrier to entry towards a culture that values holistic student development and mentorship opportunities. Creating environments that uphold diverse perspectives, encourage collaboration across student groups, and make historically underrepresented groups feel a greater sense of belonging starts with providing faculty the opportunities to develop and understand the reasons for change. A cultural change in the way a program considers what it can do to benefit its students stands to improve retention and success rates for minority students within STEM programs.

As the implementation of these findings go, translating growth of STEM programs for many universities involves making use of emerging technologies to better utilize limited resources. Where programs such as MemphiSTEP allowed for extensive mentorship and collaborative programs at universities that had adequate resources, smaller institutions could potentially struggle to produce similar benefits, especially those found in more ethnically diverse states like Texas. Technology-driven solutions such as virtual reality lab sessions or AI-driven tutoring opportunities could be used to bridge the gap between smaller and larger universities and provide higher-quality education. Additional efforts could also be directed at combining STEM pursuits with additional departments, such as social sciences or humanities, to create new opportunities for collaboration across programs and potentially attract additional students to STEM. Future work should also be aimed at addressing and identifying the most scalable components of successful STEM programs. While it is currently possible to identify the contributing factors to a successful STEM education, translating those factors so that they apply to a wider number of universities is another task altogether that's equally important. By building upon continued efforts, future directions can drive meaningful progress toward equity in STEM education. Such efforts will ensure that STEM fields accurately reflect the overall diversity of the population as a whole, as well as create innovation and excellence for every student engaging in STEM curriculum.

Acknowledge

This work is supported in part by the National Science Foundation (NSF) under Grant No. 2301868 and the National Institute of Food and Agriculture (USDA NIFA) Grant No. 2023-70020-40570 .

References

[1] S. R. Howe, "Culture at work: A comparative analysis of advertising for new york fashion week and paris fashion week," Ph.D. dissertation, University of Oregon, 2020.

[2] *Dramatic changes in science engineering landscape call for new u.s. strategies*, 2025. [Online]. Available: https://www.nsf.gov/nsb/news/news_summ.jsp?cntn_id=309993&org=NSB&from=news.

[3] S. Deitz, *The state of u.s. science and engineering 2024 | nsf - national science foundation*, 2024. [Online]. Available: https://ncses.nsf.gov/pubs/nsb20243/talent-u-s-and-global-stem-education-and-labor-force.

[4] D.-M. Institute, *taking charge: manufacturers support growth with active workforce strategies*. 2024. [Online]. Available: https://themanufacturinginstitute.org/wp-content/uploads/2024/04/Digital_Skills_Report_April_2024.pdf.

[5] *Us manufacturing could need as many as 3.8 million new employees by 2033, according to deloitte and the manufacturing institute*$_2$024, Apr. 2024. [Online]. Available: https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/us-manufacturing-could-need-new-employees-by-2033.html.

[6] S. Deitz, *The state of u.s. science and engineering 2024 | nsf - national science foundation*, 2024. [Online]. Available: https://ncses.nsf.gov/pubs/nsb20243/key-takeaways.

[7] Bpcnet.org, *Research & data – bpcnet*$_2$023, 2023. [Online]. Available: https://bpcnet.org/research-data/.
[8] C. R. A. C. for Evaluating the Research Pipeline, *Statistics & data hub, ccd: K-12 enrollment*, 2024. [Online]. Available: https://cerp.shinyapps.io/CCD-K12-enrollment/.
[9] B. of Labor Statistics, *college enrollment and work activity of 2011 high school graduates*. 2012. [Online]. Available: https://www.bls.gov/news.release/archives/hsgec_04192012.pdf.
[10] National Science Foundation, *Report-s&e indicators 2018*. [Online]. Available: https://nsf.gov/statistics/2018/nsb20181/report/sections/higher-education-in-science-and-engineering/undergraduate-education-enrollment-and-degrees-in-the-united-states.
[11] A. Bonfert and D. Wadhwa, *Tracing global trends in education | world bank gender data portal*$_2$024, 2024. [Online]. Available: https://genderdata.worldbank.org/en/data-stories/a-tale-of-old-and-new-gender-gaps#:~:text=Gender%20disparities%20in%20STEM%20tertiary,(Brixi%20et%20al%202023).
[12] K. Eagan, S. Hurtado, and M. Chang, "What matters in stem: Institutional contexts that influence stem bachelor's degree completion rates," in *annual meeting of the Association for the Study of Higher Education, Indianapolis, IN*, 2010.
[13] M. K. Eagan Jr, S. Hurtado, M. J. Chang, G. A. Garcia, F. A. Herrera, and J. C. Garibay, "Making a difference in science education: The impact of undergraduate research programs," *American educational research journal*, vol. 50, no. 4, pp. 683–713, 2013.
[14] S. Biswas, R. Benabentos, E. Brewe, *et al.*, "Institutionalizing evidence-based stem reform through faculty professional development and support structures," *International journal of STEM education*, vol. 9, no. 1, p. 36, 2022.
[15] D. B. Taylor and B. Govan, "Stem outreach in northern queensland: The importance of providing professional development and networking opportunities to educators," *International Journal of Innovation in Science and Mathematics Education*, vol. 25, no. 5, 2017.
[16] E. Borda, E. Schumacher, D. Hanley, *et al.*, "Initial implementation of active learning strategies in large, lecture stem courses: Lessons learned from a multi-institutional, interdisciplinary stem faculty development program," *International Journal of STEM Education*, vol. 7, no. 1, pp. 1–18, 2020.
[17] C. T. Belser, D. J. Prescod, A. P. Daire, M. A. Dagley, and C. Y. Young, "Predicting undergraduate student retention in stem majors based on career development factors," *The Career Development Quarterly*, vol. 65, no. 1, pp. 88–93, 2017.
[18] C. T. Belser, M. Shillingford, A. P. Daire, D. J. Prescod, and M. A. Dagley, "Factors influencing undergraduate student retention in stem majors: Career development, math ability, and demographics.," *Professional Counselor*, vol. 8, no. 3, pp. 262–276, 2018.
[19] E. B. Van Oosten, K. Buse, and D. Bilimoria, "The leadership lab for women: Advancing and retaining women in stem through professional development," *Frontiers in Psychology*, vol. 8, p. 2138, 2017.
[20] A. Elzawawy, "Increasing student enrollment and achievement in engineering and engineering technology," *American Scoeity for Engineering Education*, 2018.

[21] N. A. Bowman and K. Culver, "When do honors programs make the grade? conditional effects on college satisfaction, achievement, retention, and graduation," *Research in Higher Education*, vol. 59, pp. 249–272, 2018.

[22] K. Smith Burton, E. White, and P. Gillespie, "Fostering a stem learning community to promote student interest in stem disciplines," *American International Journal of Social Science*, vol. 5, no. 3, pp. 95–108, 2016.

[23] A. Windsor, A. Bargagliotti, R. Best, *et al.*, "Increasing retention in stem: Results from a stem talent expansion program at the university of memphis," *Journal of STEM education*, vol. 16, no. 2, 2015.

[24] H. A. Maddah, "Proposed strategic plan to improve student retention and enrollment at universities: A perspective study.," *International Society for Technology, Education, and Science*, 2021.

[25] J. K. Piper and D. Krehbiel, "Increasing stem enrollment using targeted scholarships and an interdisciplinary seminar for first-and second-year college students," *Journal of STEM Education*, vol. 16, no. 4, pp. 36–43, 2015.

[26] E. Orozco, *Forecasting Retention Among At-Risk College Freshmen Students Through Completion of a STEM Intervention*. The University of Texas Rio Grande Valley, 2023.

[27] M. B. Lee, *An Observation of the Effects of an Advisement Program, on Retention, Enrollment, and Success in Upper-Level Stem Courses*. Delta State University, 2023.

[28] M. Greenacre, P. J. Groenen, T. Hastie, A. I. d'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," *Nature Reviews Methods Primers*, vol. 2, no. 1, p. 100, 2022.

[29] K. P. Sinaga and M.-S. Yang, "Unsupervised k-means clustering algorithm," *IEEE access*, vol. 8, pp. 80 716–80 727, 2020.

[30] X. Su, X. Yan, and C.-L. Tsai, "Linear regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 3, pp. 275–294, 2012.

[31] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 12, pp. 6999–7019, 2021.

[32] D.-C. Feng, Z.-T. Liu, X.-D. Wang, *et al.*, "Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach," *Construction and Building Materials*, vol. 230, p. 117 000, 2020.

[33] Y.-Y. Song and L. Ying, "Decision tree methods: Applications for classification and prediction," *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130, 2015.

[34] P. F. Smith, S. Ganesh, and P. Liu, "A comparison of random forest regression and multiple linear regression for prediction in neuroscience," *Journal of neuroscience methods*, vol. 220, no. 1, pp. 85–91, 2013.

[35] S. Lathuilière, P. Mesejo, X. Alameda-Pineda, and R. Horaud, "A comprehensive analysis of deep regression," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 9, pp. 2065–2081, 2019.

[36] G. Van Houdt, C. Mosquera, and G. Nápoles, "A review on the long short-term memory model," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5929–5955, 2020.

[37] J. Huang and S. Kaewunruen, "Forecasting energy consumption of a public building using transformer and support vector regression," *Energies*, vol. 16, no. 2, p. 966, 2023.

[38] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation," *Peerj computer science*, vol. 7, e623, 2021.

[39] A. Crain and K. Webber, "Across the urban divide: Stem pipeline engagement among non-metropolitan students," *Journal for STEM Education Research*, vol. 4, no. 2, pp. 138–172, 2021.

[40] E. McGee, J. Mitchell, and J. Cai, "Searching for a new homeland: How geography matters in the college selection and career decisions of stem phds," *Journal of STEM Education: Innovations and Research*, vol. 25, no. 3, 2024.

[41] S. Xie, "The geography of inequality in stem: Immigration, gender, and economic disparities in the united states," Ph.D. dissertation, Arizona State University, 2024.

[42] L. Ononye and S. Bong, "The study of the effectiveness of scholarship grant program on low-income engineering technology students.," *Journal of STEM Education*, vol. 18, no. 5, 2018.

[43] B. L. Castleman, B. T. Long, and Z. Mabel, "Can financial aid help to address the growing need for stem education? the effects of need-based grants on the completion of science, technology, engineering, and math courses and degrees," *Journal of Policy Analysis and Management*, vol. 37, no. 1, pp. 136–166, 2018.

[44] B. E. Rincon and C. E. George-Jackson, "Stem intervention programs: Funding practices and challenges," *Studies in Higher Education*, vol. 41, no. 3, pp. 429–444, 2016.

[45] K. Cummings, S. Laderman, J. Lee, D. Tandberg, and D. Weeden, "Investigating the impacts of state higher education appropriations and financial aid.," *State Higher Education Executive Officers*, 2021.

[46] J. M. Harackiewicz and S. J. Priniski, "Improving student outcomes in higher education: The science of targeted intervention," *Annual review of psychology*, vol. 69, no. 1, pp. 409–435, 2018.

[47] A. M. Kassaee and G. H. Rowell, "Motivationally-informed interventions for at-risk stem students," *Journal of STEM Education: Innovations and Research*, vol. 17, no. 3, 2016.