2023 Annual Conference & Exposition
Baltimore Convention Center, MD | June 25 - 28, 2023

The Harbor of Engineering
Education for 130 Years

ASEE

Paper ID #36895

Predicting Student Success in College Algebra Classes Using Machine Learning

Dr. Zeynep Akcay Ozkan, City University of New York, Queensborough Community College

Dr. Zeynep Akcay Ozkan is an Associate Professor of Mathematics at Queensborough Community College of the City University of New York. She received her PhD in Applied Mathematics from the joint program at New Jersey Institute of Technology and Rutgers Universities (2014), with concentration on Mathematical and Computational Neuroscience. She also holds an MS degree in Financial Mathematics from Florida State University (2009). Dr. Akcay Ozkan's research interests include mathematical neuroscience, math education and data science.

Yuanhong Yu, City University of New York, Queensborough Community College
Dr. Ewa Stelmach, City University of New York, Queensborough Community College

Ewa Stelmach received her Ph.D. in Mathematics Education from Columbia University in 2019, a M.Phil. from Columbia University in 2017, a M.S in Applied Mathematics from Hofstra University in 2011 and B.S. in Mathematics from Stony Brook University in 2009. She has been teaching mathematics at Queensborough Community College, CUNY since 2011. She is always looking for innovative ways to teach her classes to inspire her students and enhance their learning experience. Over the years she participated in many departmental committees to help improve students' experience. Ewa Stelmach is a co-author of the Open Resource Educational textbook for College Algebra students. She is also the administrator and author of many problems in WeBWork, a free homework platform. Her interests include college-level teaching, mathematics education, and teaching with technology.

Predicting Student Success in College Algebra Classes Using Machine Learning

Abstract

College Algebra is a gateway course for STEM majors with large enrollment and low passing rates. We analyze the factors which contribute to student success in College Algebra courses at an urban community college. Characteristics and grades of over twenty thousand students who were enrolled in College Algebra courses between the years 2017 and 2022 have been analyzed. Among the students' characteristics being studied are gender, ethnicity, age, first-generation college status, placement exam scores, grade point averages (GPA), whether they are freshmen or transfer students. The course modalities include online, hybrid or in-person. We study correlations between factors that affect student success. Using k-nearest neighbor and decision tree algorithms, we predict student success based on the student characteristics and course features. Using Chi-Square Test of Independence, we show that passing rates of students depend on gender, ethnicity, age, overall GPA and whether they are freshmen or transfer students. Passing rates also depend on the modality of the course and the semester (fall or spring) the course is taken. With both supervised machine learning algorithms used, the probability of students passing were predicted with approximately 85 percent accuracy. Our results show that machine learning models can successfully be used on student data to predict course outcomes which can enable early intervention to those students with higher chances of failure in the course. Our findings may encourage college administrations to use machine learning for predicting student success and be able to provide better advisement to incoming students regarding course selection.

Keywords: Retention, College Algebra, Student Success, Machine Learning

Introduction

Academic institutions have always cared about and searched for ways to improve student success and retention. With the recent decline in student enrollment and retention rates nationally, improving student performance and completion rates has become an important objective for institutions [1]. One of the major changes City University of New York (CUNY) undertook was to end offerings of traditional remedial courses as the research showed they hindered student progress toward their degree [2].

College Algebra is a gateway course for STEM majors at the Queensborough Community College of City University of New York (CUNY) with high enrollment but low passing rates. While there have been initiatives constantly emerging, the success of students taking core mathematics courses continue to decline [3].

Another recent change in higher education is the increased number of online course offerings. While it may be more appealing for students to take some form of an online class [4], the research shows that online classes have lower passing rates [5]. Therefore, students should be advised to carefully consider the pros and cons of different course modalities and choose the one that they would benefit most from.

Data of over twenty thousand students who have taken College Algebra courses Queensborough Community College between the years 2017 and 2022 have been analyzed. Student and course characteristics have been studied to identify factors that have correlations with student success. Two supervised machine learning models, K-Nearest Neighbors (KNN) and Decision Trees have been used to predict student success based on student and course characteristics.

Among the students' characteristics being studied are gender, ethnicity, age, first-generation college status, placement exam scores, grade point averages (GPA), mathematics course taken and grades received prior to enrollment in College Algebra. The course modalities include in-person, hybrid and online.

Literature review

Much attention has been devoted to understanding factors that affect student success. This subject has been studied from different angles and utilizing various methods. We first discuss some of the studies which analyzed student success using statistical methods.

Reyes [6] conducted a study to determine if factors such as course length (8-week, 16-week), gender, age, and ethnicity affect students' performance in algebra courses. Grades were analyzed using non-parametric Chi-Square (v2) tests of independence. Reyes' study found that some student characteristics were statistically significant while other student characteristics were not but admitted that her sample size may have been too small.

Smith Jaggars [7] studied online learning in the community college setting. Their focus was on patterns of students' online course-taking and their performance in online vs. face-to-face courses as well as the factors and student characteristics affecting online course performance. Jaggars' study shows that students usually perform worse in any given course if taken online rather than in person, but the difference in performance is more evident for some demographic groups than it is for others.

O'Connell et al. [8], examined historical student data over twenty thousand students in Introductory College Algebra. Their study shows that student success is best predicted by student past performance and experiences, GPA, and number of accumulated credit hours, with other student characteristics having a smaller impact on student performance. Notably, it was found that time spent on assignments is associated with higher grades. O'Connell et al. noted that their study involved only one university with high diversity in the student population and that the results they achieved in their study may or may not transfer to other institutions with different levels of diversity.

Ongoing improvements in predictive power of machine learning algorithms made these methods gain popularity in many application areas. Over the last few years, machine learning methods have been utilized in several studies to predict student success.

Zeineddine et al. [9] conducted a study testing the accuracy of different machine learning methods in predicting student success based on several student's characteristics. The study has shown that decision tree method was 90% accurate and k nearest neighbor method was 83% accurate. Many other methods were also studied, demonstrated various levels of accuracy depending on the method. Zeineddine et al. concluded that automated machine learning methods can be highly effective at predicting student performance and providing administration with early intervention options for higher risk students.

Yehuala [10] analyzed factors affecting student success with a dataset of around eleven thousand undergraduate students. Using the classification rule generation process based on the decision tree and Bayes, Yehuala found that gender, number of students in a class, number of courses provided in a semester and field of study are the main factors that affect student performance. The results provided constructive recommendation to class planners to structure curriculum that will maximize student success [10].

Jiao et al. [11] used genetic programming to develop artificial intelligence-enabled prediction models that analyze students' learning process and summative data from an online engineering course and obtained high accuracy on similar courses when the same model is applied. The study found that predictive model serves as a viable tool to predict the learning performance of students [11].

Kabakchieva [12] conducted a data mining study using decision tree, Bayes, and Nearest Neighbor algorithms utilizing data from over ten thousand students' pre-university characteristics at a Bulgarian University. Her study revealed that the prediction rates are not remarkable, varying between 52-67% accuracy. However, Kabakchieva admits that this was only a preliminary study and that tuning the parameters of the algorithm and data sets could still yield more promising results.

Gandy et al. [13] conducted a study using decision tree models and neural networks that utilize the strength of statistical learning to predict student success in the first two years of college. Fifty variables were examined including student retention rates as well as gender, race, academic and economic data. Results indicate that predicting student success probabilities using these methods were strong.

Methodology

Data Set

Roughly 2500 students enroll in a College Algebra section every semester at Queensborough Community College of CUNY. Data from students enrolled in the College Algebra courses between the years 2017 and 2021 have been requested from the Office of Institutional Research and Assessment. The starting term for the data set corresponds to the time when online or hybrid classes have started to be offered while the ending term is the term for which most recent data was available at the time of the data request. Data for approximately 23,000 students meeting these criteria has been obtained.

Table 1. Names and data ranges for student and course characteristics.

Student/Course Features	Data Ranges
Semester of enrollment	Fall, Spring
Year of enrollment	2017-2021
Modality	In-person, online, hybrid
Ethnicity	Hispanic, Black, White, Asian or Pacific Islander, American Indian or Native Alaskan
Age	13-73
Gender	Female, Male
New Student Description	First-time Freshmen, Transfer
GPA	0-4
Credits completed	0-188
Course Grade	A-D, F, W, Other
Math/Reading/Writing Placement Exam Score	Passed, Failed, Exempt, Not tested
First Generation Status	Yes, No

Data set comprises characteristics of students and the sections they are enrolled. The students' characteristics include age, gender, ethnicity, first-generation college status, placement exam scores, GPA, credits completed, and whether they are freshmen or transfer students. Course characteristics include year and semester of enrollment, and course modalities. The course modalities are online, hybrid and in-person. All features used are listed in Table 1 together with their ranges of values.

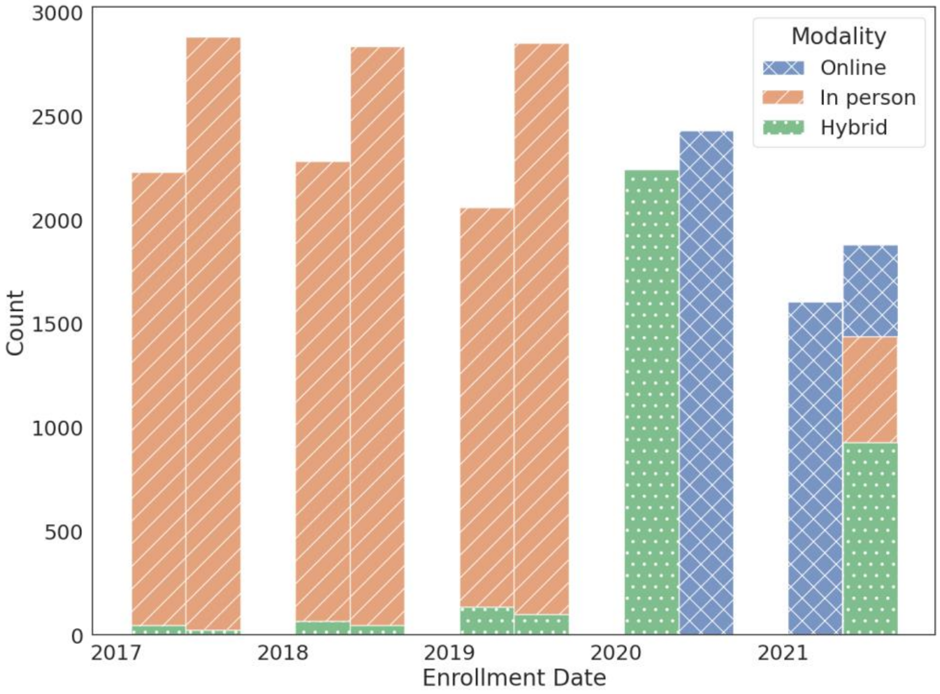

Figure 1. Number of enrolled students each semester

Figure 1. shows the number of students enrolled in each semester during the course of our study. The modality of courses prior to Spring 2020 were mostly in-person with a small percentage being hybrid. All sections that started in-person or hybrid transitioned to online in Spring 2020 semester due to the Covid-19 pandemic, therefore all sections are considered as hybrid in this semester. Fall 2020 and Spring 2021 semesters were fully online. All three modalities have been offered during Fall 2021. More students have been enrolled during Fall semesters while a slight decline in the number of total enrollment over the years is observed.

The ages of students enrolled in the course is shown in Figure 2a. The students' ages range from 16 to around 60, shown in bins of two. About half of the students' ages fall between 18 and 20. The grade distribution is shown in Figure 2b. We observe that a quarter of students have withdrawn from the course. A successful completion of the course is considered as receiving a grade of C or above. The total percentage of students completing the course with a C or above is 48% during the course of study. We refer to this rate as the passing rate.

Methods

The data processing and analysis steps include data preparation, exploratory data analysis, model creation, model evaluation and parameter tuning. An iterative approach was developed to gradually narrow down the most relevant parameters and improve model evaluation score.

The Python programming language was used for this study. NumPy and Pandas packages were mostly used for data preparation and cleaning, while Matplotlib and Seaborn were used for data visualization. Finally, Scikit-learn was used for machine learning model building and tuning.

Data Analysis

The graphics in this section demonstrate the correlations between the characteristics of the students and their success in the College Algebra course. We calculate passing rates among different groups and share these values in Table 2. We also use Chi-Square Test of Independence to show that these relations are statistically significant.

According to Figure 3a, students who have identified themselves as belonging to Hispanic or Black ethnicity groups constitute a larger part of the student body. Among these groups, the passing rates are 0.43 and 0.42, respectively. On the other hand, among the students who have identified themselves belonging to the Asian or Pacific Islander ethnicity groups, the passing rate is 0.60. The difference between passing rates of different ethnic groups are statistically significant based on the Chi-Square Test.

Another student feature that shows correlation with the grades is gender. The percentage of receivers of each grade in each gender is shown in Figure 3b. For passing grades (A, B and C), these percentages are higher in female students. The overall passing rates for female students (0.51) is significantly higher than the passing rate for male students (0.44) which is consistent with the graph.

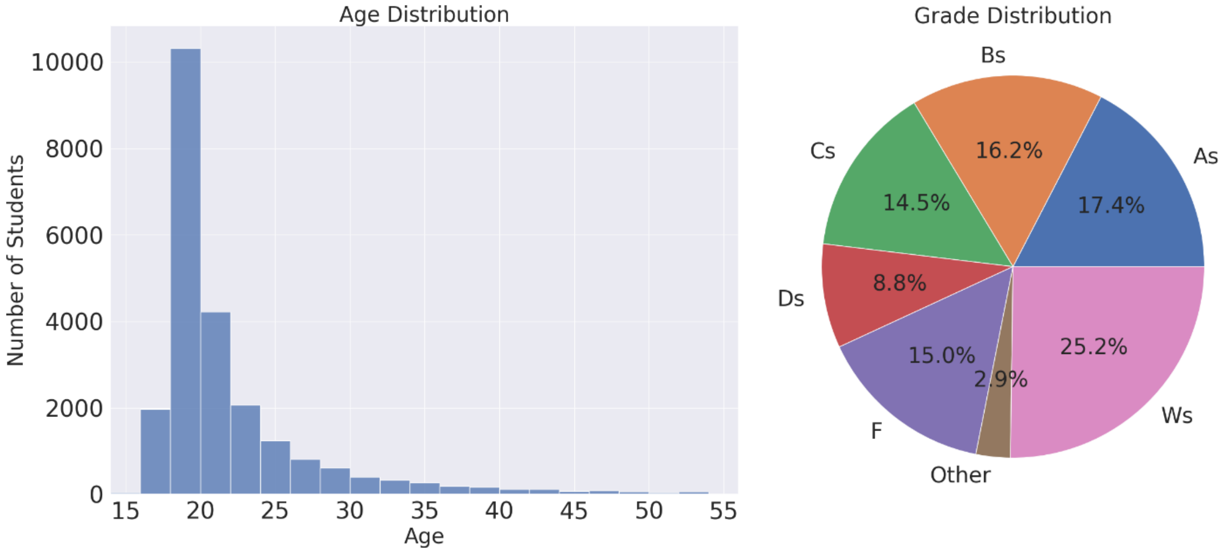

Figure 2. a) Age distribution **b)** Grade distribution of College Algebra students

Figure 4a demonstrates the positive correlation between ages of students and their College Algebra grades. The course grade average is calculated by converting the letter grades to a numeric value (Grades A-F ranging between 4-0) and is plotted on the vertical axis, while the ages are shown on the horizontal axis in bins of two. An increase in the average student grades is observed along with the increase in age groups. Chi-Square test confirms this positive correlation.

Table 2. Passing Rates based on student or course features.

Student/Course Feature	**Values**	**Passing Rates**
Semester of enrollment	Fall	0.46
	Spring	0.50
Modality	In-person	0.45
	Hybrid	0.48
	Online	0.57
Ethnicity	Black	0.42
	Hispanic	0.43
	American Indian or Native Alaskan	0.45
	White	0.52
	Asian or Pacific Islander	0.60
Gender	Male	0.44
	Female	0.51
New Student Description	First-Time Freshmen	0.48
	Transfer	0.58

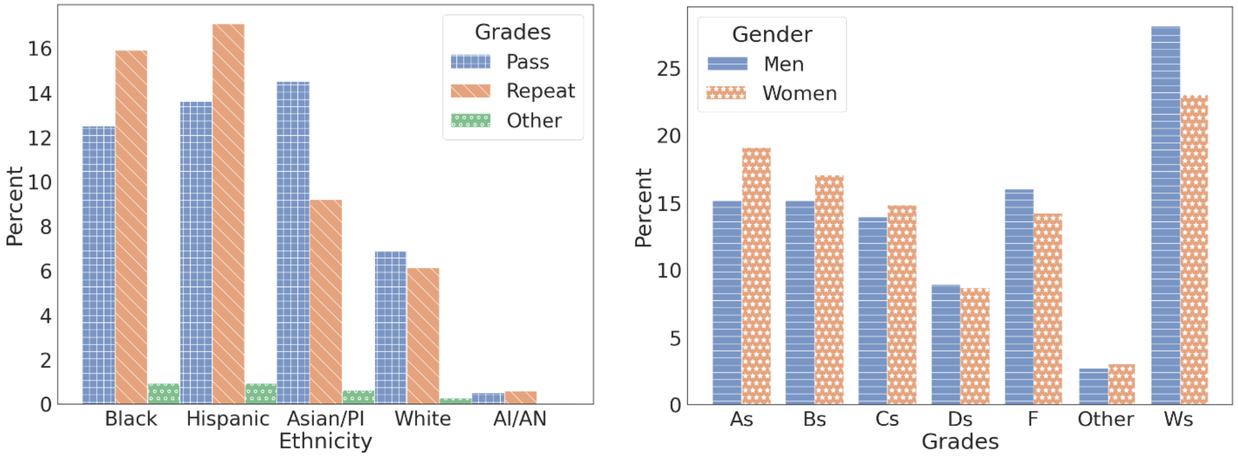

Figure 3. The relation between **a)** pass/fail rates and ethnicity **b)** grades and gender of students

A positive correlation is observed between students' GPA and College Algebra grades as shown in Figure 4b via box plot. This positive correlation is an expected relation due to confounding factors.

Another student feature that showed dependence with course grade is 'New Student Description' which categorizes students as first time freshmen or transfer. Passing rate for first-time freshmen students was 0.48 while it was 0.58 for transfer students.

The modality of the course also had a correlation with the student's passing rates. The passing rates for in-person sections was 0.45, while it was 0.48 for hybrid and 0.57 for online sections. The higher rate for online sections could possibly be explained by flexibility provided to students during the pandemic or by difficulty in proctoring exams online.

Lastly, the passing rates depended on the semester of enrollment in the course. The average passing rate was 0.46 for College Algebra sections taken in Fall semesters while it was 0.50 for sections taken in Spring semesters. This dependence is also found to be statistically significant by Chi-Square Test.

We also looked for any dependency between grades and entrance exam scores on mathematics, reading and writing. The average College Algebra passing rate for students who were categorized as 'not tested' was the highest at around 0.60 in each of the three subjects. Students who 'passed' the initial math exam had a higher passing rate (0.51) in College Algebra compared to students who 'failed' this exam (0.42). The passing rates did not differ significantly among students who passed, failed or who were exempt in reading and writing entrance exams, and was around 0.46 for all three categories. Another student feature which did not show significant dependency with course grades was first generation college status of the students.

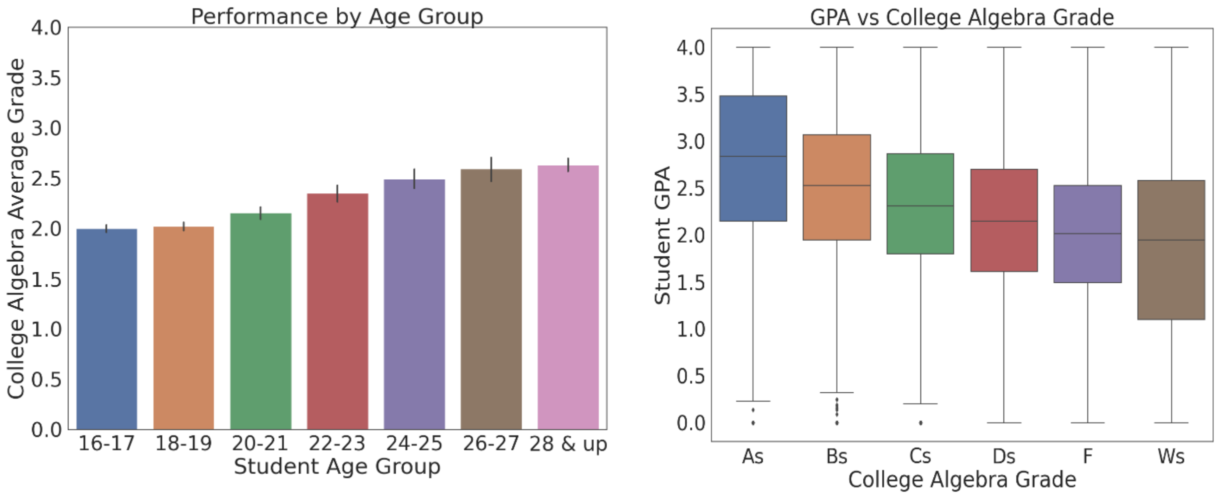

Figure 4. The relation between **a)** ages, **b)** GPAs of students and their College Algebra grades

Predictive Models

This study employed both KNN and Decision Tree algorithms. Hyper parameters were adjusted to attain the highest F1 and Jaccard scores. Described below are the data cleaning, pre-processing and training steps for creating the machine learning models.

Data Cleaning

The quality and format of the data play an important role in the performance of machine learning algorithms. A series of processes were performed on the original student data set before it was fed into the model for evaluation.

The data set that was obtained initially received some features that included many missing data points. Those features were not included in the descriptive analysis of the data or in creating predictive models. A small percentage of course grades were labeled as ‘Other’ which could not be identified as passing, failing or withdraw. These data points were included in the descriptive analysis but not in the predictive models.

Multiple entries for some students exist in the data set due to those students repeating the College Algebra course between the years 2017 and 2021. The number of attempts each student made to complete the Algebra Courses were calculated and served as one of the input parameters of the model. Only the grade obtained in their most recent attempt was utilized for evaluating the model.

Additionally, student records with grades that are not included in the computation of the GPA, referred to as other grades (such as incomplete, non-credit, transfer, etc.) are removed from the data set.

Table 3. Evaluation of the machine learning algorithms used.

	K-Nearest Neighbors		**Decision Trees**	
	Train	Test	Train	Test
Accuracy	0.854	0.842	0.861	0.849
Precision	0.867	0.866	0.875	0.874
Recall	0.961	0.947	0.959	0.945
F1 Score	0.912	0.905	0.915	0.908
Jaccard Score	0.837	0.826	0.844	0.832

Preprocessing

The next step involved transforming the data to enable the algorithm to handle non-numerical information and prevent biases. Categorical variables, such as semester of enrollment, gender, ethnicity, course grade and placement exam results, were transformed into numerical representations so the algorithm could process the information. The course modality was also transformed through one-hot encoding.

After all the data was in numerical form, it was normalized to ensure that all attributes fell within the range of 0 to 1. This normalization is particularly helpful in avoiding biases in the KNN algorithm when calculating distance between data points.

Training

Two-thirds of the data set were used for the training of the models. Various hyper parameter combinations were tested for both the KNN and Decision Tree algorithms to attain the highest accuracy.

The hyper parameters tuned for KNN included the number of nearest neighbors, which determines how many neighbors to consider when making prediction on a specific data point. Another hyper parameter tuned was the distance metric, which determines how to calculate the distance between data points and the algorithm used such as ball tree or K-D tree.

As for the decision tree, the hyper parameters that were modified were the maximum depth, which defines the maximum level the tree can reach, the criterion, which determines the method to evaluate the quality of the split, the minimum sample split, and minimum sample leaf, which are guidelines that specify the minimum number of samples required to split a node or be a leaf node. All models were evaluated through cross-validation and the combination with the highest accuracy for each model was evaluated against testing data set for final evaluation.

Results

Both KNN and Decision Tree models produced accurate predictions. F1 scores were found to be 0.91 while the Jaccard indices were found to be 0.83 for both models. The evaluation metrics for each algorithm are given in Table 3.

The analysis shows that most of the student and course characteristics analyzed display correlation with the College Algebra course outcomes. Specifically, female students, students of Asian or Pacific Islander origin, older students, students enroll in online sections have a higher chance of performing better compared to their peers. A positive correlation between students' performance in College Algebra and GPA was also observed. Transfer students tend to perform better than first-time freshmen. Lastly, students performed slightly better during Spring semesters. Although the differences between passing rates in different categories are not very high, the cumulative effect of these features has a strong predictive power.

Conclusions and Future Directions

This study analyzed two machine learning algorithms, KNN and decision tree, to predict outcomes in College Algebra courses based on characteristics of students and courses. Correlation analysis showed that student characteristics such as age, gender, ethnicity, GPA, new student description, and course characteristics such as semester of enrollment and modality of the course all had correlations with the grades.

The dependency of course success to student and course features were shown to be statistically significant and their cumulative effect was sufficiently strong for the machine learning algorithms to predict the course outcomes with 85 percent accuracy. We conclude that machine learning algorithms, specifically KNN and decision tree algorithms, can be used to predict course outcomes based on the students' demographics.

The results of this study can be used by college administrations to make predictions about student success in College Algebra courses based on student characteristics. This can also be used by administrations to create new interventions that better support students in their success and reduce student failure and better advisement can be provided to the incoming students regarding the course selection.

The elegance of using machine learning to analyze student performance and predict future student performance based on student characteristics is that the process can be automated to a large extent at any particular institution. Instead of conducting an exhaustive and time-consuming study with large data sets manually each year, machine learning allows researchers to quickly enter large data sets into the software for fast and efficient analysis. Yearly or semi-annual data injections into the software will help in keeping the recommendations and interventions produced by the software up to date, based on the ever-changing student body and their associated data. A well-tuned machine learning software with frequent student data updates can prove to be a powerful tool for institutions.

As future directions, other machine learning methods such as Logistic Regression or Support Vector Machines may be used. Further characteristics such as course history of students may be utilized to improve the predictive power of the machine learning algorithms.

References

[1] National Student Clearing House Research Center, “Stay Informed with the Latest Enrollment Information.” https://nscresearchcenter.org/stay-informed/.

[2] City University of New York, “CUNY Ends Traditional Remedial Courses,” 2023. https://www1.cuny.edu/mu/forum/2023/01/12/cuny-ends-traditional-remedial-courses/.

[3] The Office of Institutional Research and Assessment, “Queensborough Community College Fact Book,” 2021. [Online]. Available: https://www.qcc.cuny.edu/oira/docs/Factbook-2021.pdf.

[4] J. Robert, “2022 Students and Technology Report: Rebalancing the Student Experience.” https://www.educause.edu/ecar/research-publications/2022/students-and-technology-report-rebalancing-the-student-experience/modality-preferences.

[5] C. C. R. C. Columbia University, “Community College FAQs.” https://ccrc.tc.columbia.edu/community-college-faqs.html.

[6] C. Reyes, “Success in Algebra Among Community College Students,” *Community Coll. J. Res. Pract.*, vol. 34, no. 3, 2010, doi: 10.1080/10668920802505538.

[7] S. S. Jaggars, *Handbook of Distance Education*. New York, 2018.

[8] K. A. O’Connell, E. Wostl, M. Crosslin, T. L. Berry, and J. P. Grover, “Student Ability Best Predicts Final Grade in a College Algebra Course,” 2018, doi: https://doi.org/10.18608/jla.2018.53.11.

[9] H. Zeineddine, U. Braendle, and A. Farah, “Enhancing prediction of student success: Automated machine learning approach,” *Comput. Electr. Eng.*, vol. 89, p. 106903, Jan. 2021, doi: 10.1016/J.COMPELECENG.2020.106903.

[10] M. A. Yehuala, “Application Of Data Mining Techniques For Student Success And Failure Prediction,” *Int. J. Sci. Technol. Res.*, vol. 4, no. 5, 2015.

[11] P. Jiao, F. Ouyang, Q. Zhang, and A. A.H, “Artificial intelligence enabled prediction model of student,” *Artif. Intell. Rev.*, vol. 55, pp. 6321–6344, 2022, doi: https://doi.org/10.1007/s10462-022-10155-y.

[12] D. Kabakchieva, “Predicting Student Performance by Using Data Mining Methods for Classification,” *Cybern. Inf. Technol.*, vol. 13, no. 1, pp. 61–72, Mar. 2013, doi: 10.2478/cait-2013-0006.

[13] R. Gandy, D. Kasper, and A. Luna, “Creating a Student Success Predictor Using Statistical Learning.” [Online]. Available: https://www.apsu.edu/dsir/reports/creating_a_student_success_predictor.pdf.