Worcester State University

MA–470: Capstone Experience

Fall 2022

Predicting Student Performance Using Data Mining Techniques

Author: Avery Oldakowski

Abstract

The aim of this study is to determine whether we can build an accurate model to predict the cumulative GPA of Worcester State University students using pre-admission and collegiate demographic and academic factors. Using data mining techniques such as linear regression, PART, OneR, ZeroR, J48, REPTree, Random Tree, Random Forest, K-Nearest Neighbors, and Naïve Bayes algorithms, we determined what factors have the most statistically significant contributions to cumulative GPA. We found that certain majors and years' GPAs contribute the most, along with graduation and return rates. Race can also be a contributing factor in cumulative GPA. Our linear regression model was the most accurate in predicting students' academic performance. Being able to examine which predictive factors increase or decrease GPA, administration and faculty can provide support where it is most needed.

Introduction

Understanding what majors, demographic backgrounds, and specific classes contribute to a student's final grade, whether it be a final course grade or an overall GPA, is pertinent to finding the best way to educate students. Identifying weaknesses in the education system can lead to significant and necessary improvements that can guarantee an improved education for students of all backgrounds. Many studies cited below found that, on the individual course level, midterm grades are the earliest classifiers for a student's final course grade. Knowing in advance which students will be at risk of failing is extremely important so that educators will be able to intervene with the needed support which can potentially restore the students' grades.

Recognizing which disciplines produce lower and higher GPAs can identify systemic bias towards or against certain majors. Identifying these facts may not alter them, but may give administration, faculty, and students the motive and facts to modify the system and create a better educational experience.

The purpose of this study is to determine whether the cumulative GPA of a student can be accurately modeled using academic and demographic data from existing students. We aim to answer the following questions:

1. Can we build accurate models to predict students' cumulative GPA both numerically and by category?

2. What factors most accurately predict students' final cumulative GPAs?

By answering these questions we can understand what factors put students at risk of doing poorly in their academic career. With this information we can alert students, administration, advisors, and faculty to warning signs of failure or inadequate grades, and give students the warning, support, and resources that they may need to improve their grades and be successful in college.

Literature

There have been many similar studies conducted to determine what factors most contribute to student success and, conversely, student lack of success in practically every discipline. Educators, administrators, and students alike want to know *why* grades are what they are, and what they can do to improve them.

Demographic and Academic Features in a Student Performance Prediction

Bilal et al. [1] analyzed students' demographics, pre-admission academic achievements, and performance in their first semester to predict their performance in their last semester in a Doctor of Veterinary Medicine program in a Pakistan university. Instead of predicting a numerical value for GPA based on these factors, the authors assigned students into two different achievement groups; students with a GPA of 3.0 or higher were classified as high-performing and students with a GPA of less than 3.0 were classified as low-performing. The author then explained that, for their study, the data was imbalanced, meaning there was a lot of data in the high-performing class. This skews any model used because suppose, for example, 99 percent of data points are high-performing. Then any model could predict, based solely on the distribution, that any given student would be high-performing with 99 percent accuracy. To combat this imbalance, the authors used the Synthetic Minority Oversampling Technique (SMOTE). This creates synthetic samples of the minority class data points so that there is an even number of points of each trait. The data can then be used to create an accurate model that does not depend on the distribution of traits.

The authors used machine learning algorithms such as decision trees, Random Forest, Support Vector Machine, K-Nearest Neighbors, and logistic regression to predict students' final semester grades, then chose the model with the highest accuracy as their final predictive model. Using 15-fold cross-validation with 85% of the data used to create the model and the remaining 15% of the data used to test the model, the authors used a Support Vector Machine algorithm to build their model. This SVM model had a 93% accuracy, the highest of the 5 machine learning methods. To determine which factors most contributed to the students' scores, the authors used decision trees to examine the factors' contributions. The results of the study show that students' grades in Biology, English, Islamiat, and Urdu courses contributed most to student performance.

The rules that were determined from the decision trees are as follows: Students with grades greater than 69% in Biology and grades greater than 91% in Islamiat fell into the high-performing group. Students with grades less than or equal to 58% in English fell into the low-performing group for final semester GPA. This research found that demographics did not play a significant role in predicting academic performance in Doctor of Veterinary Medicine students at this university. The study acknowledged that other studies have found that demographics did play a significant role in predicting student success. This, however, could be due to geographic location, subject, program of study, native language, or other various factors. These findings can be used to support students in Urdu, Islamiat, Biology, and English so as to increase students' final semester GPAs and better understand student academic performance.

Determining what courses most contribute to students' overall success or failure is important: it gives educators the opportunity to provide more support for students in critical courses and to identify at-risk students without analyzing performance in every course. Similarly, by looking at student data on a broader level as opposed to course-level, potentially at-risk students can be identified earlier on in their collegiate journey. Identifying critical courses in the STEM field in particular is extremely helpful for students, administrators, and professors to know which courses will require more support for students.

STEM Courses are Harder: Inter-course Grading Disparities with a Calibrated GPA Model

Tomkin and West (2022) [6] conducted an experiment on 64,860 students from the College of Engineering and the College of Liberal Arts and Sciences at the University of Illinois over a ten-year period. Their overall aim was to determine whether there was a grade disparity between STEM and non-STEM majors. They found that GPA in its current form (adding up 4.0-scaled GPA from each class and dividing by the number of classes) is not the most accurate measure of students' success. It may be difficult to determine whether a STEM major is more difficult than a non-STEM major solely based on observed GPA, since students with higher academic ability may choose to pursue a more rigorous field in STEM, so their grades would reflect a lower grade from taking classes that are on average more difficult than non-STEM courses. The authors' solution is to build an unbiased model to measure student ability by predicting what grades students would get in every course at an institution. This would give the authors an accurate basis for comparing STEM courses and non-STEM courses, since observed GPA is already biased and can therefore not be used as a tool for comparison. They aimed to find a logistic model of GPA that would better represent students' academic performance based on their abilities and the difficulty of the classes they have taken. Using this model, they determined that courses required for STEM majors are graded more harshly than courses for non-STEM majors.

They also aimed to determine to what extent standardized test scores predict academic performance. They found that the Pearson correlation (a coefficient between -1 and 1, -1 being a perfect negative correlation and 1 being a perfect positive correlation) between ACT scores and observed GPA is $r = 0.20$ and $r = 0.49$ for between ACT scores and calibrated GPA. This means that ACT scores are a better prediction for calibrated GPA, but still show some insight into observed GPA. Tomkin and West also found that, using the calibrated GPA model, there was no gender difference in GPA for STEM students. Based on observed GPA, however, there was evidence that women in STEM had lower GPAs than men in STEM.

Recognizing that STEM courses are graded more stringently than courses of other majors is an important connection to make. This gives us potential insight into any predictive model that factors in major or course type. By categorizing certain majors into a STEM group, we can determine what effect these majors may have on overall GPA and success or failure academically.

Education Data Mining: Prediction of Students' Academic Performance using Machine Learning Algorithms

In a 2022 study on educational data mining, Yağcı examines the effect of midterm grades on students' final semester grades [7]. Using a set of 1854 students from a state university in Turkey, Kırşehir Ahi Evran University, data including faculty, department, midterm grades, and final grades were recorded and used to build models to predict what a student's final grade in a Turkish Language I course would be at the end of the 2019-2020 fall semester. The goal of the study was to build models and compare their accuracies in order to make predictions about students' final grades using midterm grades. This has the potential to alert students and faculty of students who are more likely to do poorly in the class or fail. Because midterms were 9 weeks before finals, using these grades to make predictions gives students enough time to change their study habits and make improvements on their academic practices. Random forests, neural network, linear regression, support vector machine, and k-nearest neighbor algorithms were used to create the models in this study.

The models in this study were used to predict which grade category each student fell into. These ranges are as follows: Category 1 - [0, 32.5), Category 2 - [32.5, 55), Category 3 - [55, 77.5), Category 4 - [77.5, 100]. Yağcı found that neural network and random forest had the highest accuracy of 74.6%, although all methods had accuracies ranging from 70-75%, meaning 70-75% of data points were classified into the correct categories using these methods. The results of this study show us that by using predictive analytics in educational data mining, students will have the opportunity to correct their methods and form new habits to improve their chances of success. By using accurate models such as these, students and faculty will know when students have a higher probability of failing or doing poorly, and the support they require to do well can be obtained to produce a higher grade than predicted by the model.

By discovering the earliest point in the semester that final semester grades can be predicted, one is able to identify students who may need more support in that semester. By broadening predictive models to identify the earliest point in a student's entire collegiate career, we can recognize which students are at a higher risk of failing and give them the resources and support they need to be successful.

Artificial Intelligence Based Model for Prediction of Students' Performance

Stadlman et al., in a 2022 study of student performance, utilized data from 133 undergraduate Mechanical Engineering students from Rowan University in New Jersey from the 2020-2021 school year to analyze the effects of online learning in STEM fields [5]. They determined that students' final course grades can be predicted to fit into categorical grade values (Excellent, Good, Fair, Failure) based on artificial intelligent data mining methods. Because of how small the data set was, the authors were unable to predict a numerical value for final term grades. The authors gathered student retention information by having students fill out a form answering questions about the lecture each class and about each new topic within the lecture.

Using logistic regression, random forest, support vector machine, k-nearest neighbors, decision tree, and ensemble learning, they determined that student retention in the first and third quarters of a class session have a strong correlation with end of semester final grades. They also found that midterm grades are the earliest predictors of students' final grades. This can give students a heads up based on early warning signs of failure so they have plenty of time to recover their grades before the end of the semester. Unfortunately, although the authors were able to accurately able to predict final grade categories based on their data, the data set was too small to be able to predict numerical values and there were also no instances where students failed the class, therefore there was very little data to accurately categorize students at risk of failure. This limits the scope of the prediction model, because the aim of this study is to identify students who are at risk of failing or getting lower grades in the course.

Methods

Data for this study was collected from the Worcester State University Office of Assessment and Planning. Student information from 1816 students was gathered from students whose start years are 2014, 2015, and 2016. From these students, first-year admissions information such as demographics, high school GPA, SAT and ACT scores, and the city where the students are from was collected. Demographic information includes gender, race/ethnicity, ALANA/BIPOC, and whether or not they were a first-generation student. Along with this, data from throughout the students' time at the university was also recorded. This includes the start year, whether the student was a resident or commuter, GPA from each year, cumulative GPA, whether the student returned each year, major that the student applied and was accepted as (no change of major was recorded), whether they graduated, whether they graduated in four years, whether they graduated in 6 years, and whether they graduated in 8 years (note that 8 years is the period from the earliest start year recorded in the data to the time that this data was gathered and analyzed - if the student has not graduated within 8 years, they have not graduated since they began at Worcester State University). Each student was assigned a random ID number so student information remains anonymous.

Dataset

This data set is composed of students who are 62.11% female and 37.89% male, 31.66% first-generation students and 33.81% continuing generation students, and students who are 71.97% white and 23.07% ALANA/BIPOC including 5.34% black or African American and 0.22% Native American or Alaska Natives. 94.38% of students were from Massachusetts and

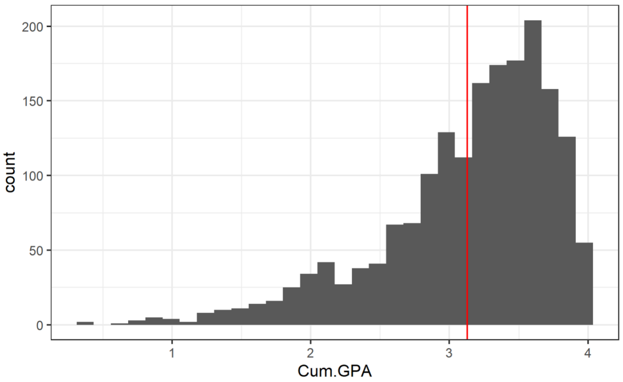


Figure 1: Histogram of cumulative GPA values

58.04% of students were from Worcester County. Only 7.93% of students submitted ACT scores, while 91.49% of students submitted SAT scores to the university. Of the students included in the study, 38.99% of students were commuters and the remaining 61.01% of students were on-campus residents for the majority of their time at Worcester State University. Figure 1 shows us the distribution of cumulative GPA within the data set, which is left-skewed with an average of 3.13. Figure 2 shows us the distribution of each year's GPA, which are similarly left-skewed and with similar averages around 3.10. Figure 3 shows us the distribution of majors across each student in the data set, the major with the most students (besides Undeclared) being Business Administration with 201 out of 1816 students.

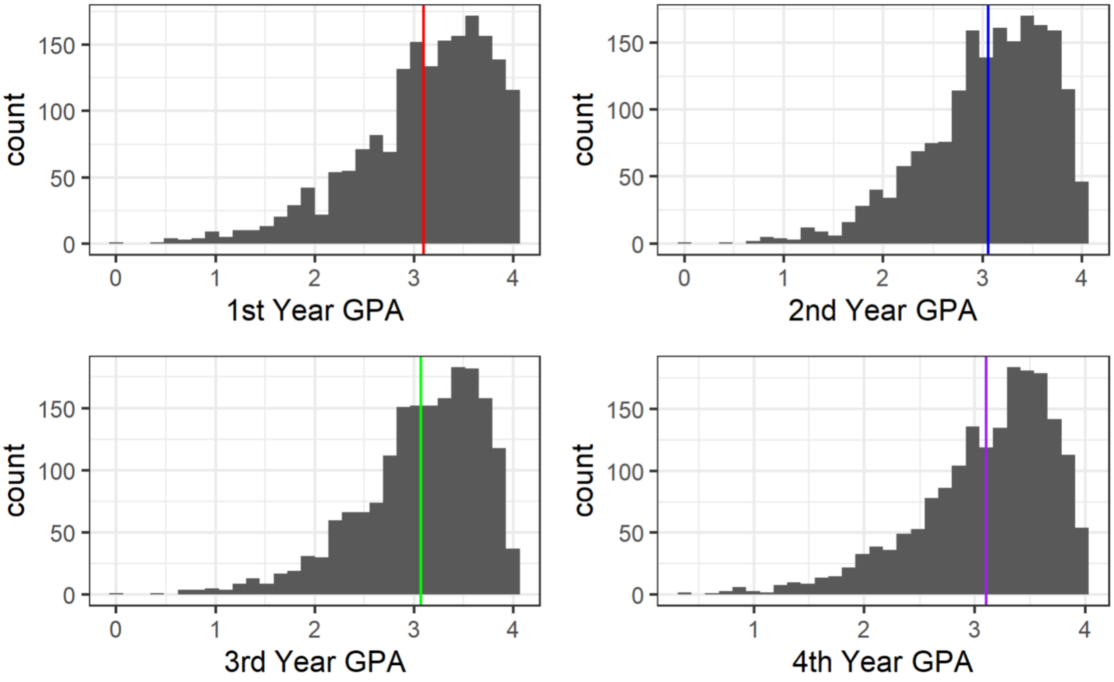


Figure 2: Histograms of each year's GPA values

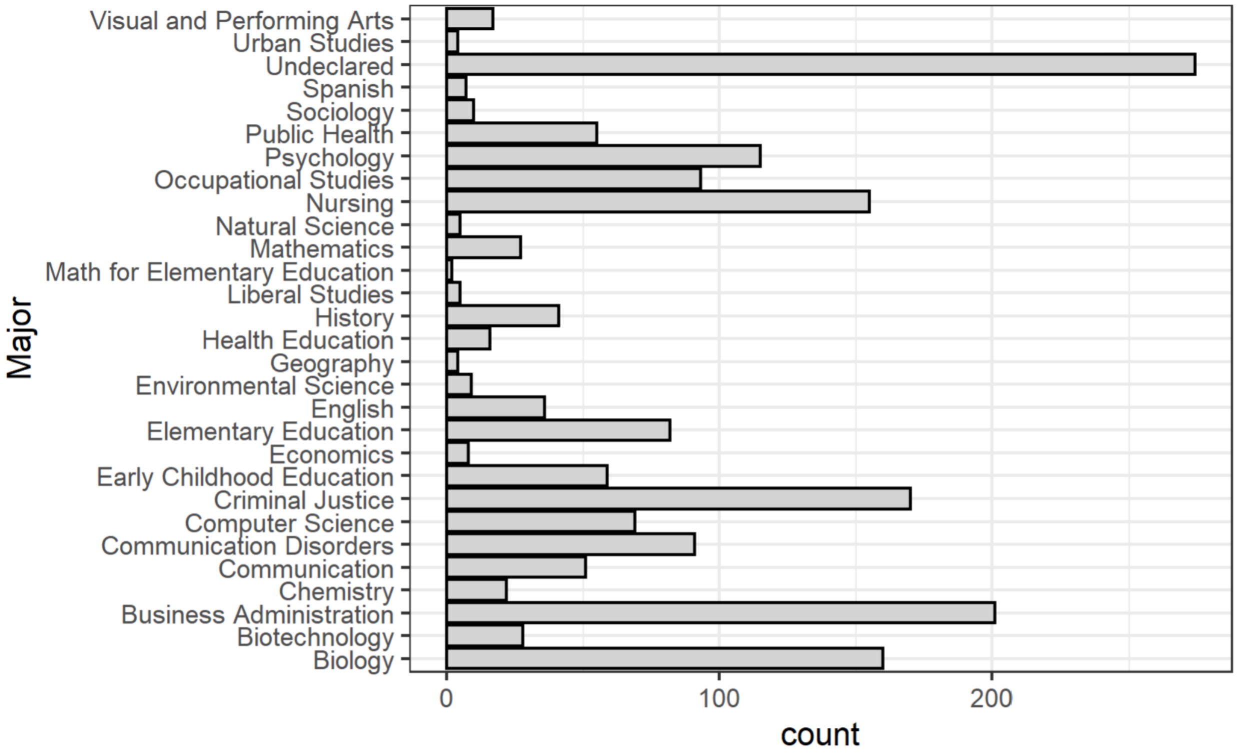


Figure 3: Histogram of each students' major at time of application and acceptance

Data preprocessing

The goal of this project is to accurately predict a student's cumulative GPA based on different academic and demographic factors. Before creating a model, however, the data must be 'tidied', meaning it must be altered so that there are no error messages. An error message can appear when there are any missing values in a row or column in a set of data. For example, 'SAT Score' is one of the attributes in the data set, however, since SAT scores are not required to be submitted for acceptance at Worcester State, many students are missing this value. Computer programs have a hard time understanding what to do when there are missing values in a data set, so one must decide what to do: either delete the entire column/attribute and ignore SAT scores as a factor all together, give the students with missing SAT scores the average score from the scores that aren't missing, delete all of the students who didn't report their scores, turn the attribute into a 'zero-or-one' attribute or 'indicator variable' that only tells whether or not a student reported their SAT scores at all and doesn't show what the score actually is, or use another strategy to get rid of the missing values. For all of the attributes with missing values, one must make those decisions before proceeding to visualize, analyze, and model the data without error.

Many variables were already free from missing values such as ID, gender, race/ethnicity, ALANA/BIPOC, first generation, and housing. However, some variables required altering. Some values in the 'State' attribute were missing, and upon further investigation it was found that these students were international, so the missing value was replaced with "Non

US". Similarly for the "County" attribute, missing values were replaced with "Non MA". For the "HS GPA", and "SAT Total" attributes, missing values were replaced with the mean of all of the recorded values. Since very few students reported ACT scores when applying to the university, the majority of students were missing this value, thus replacing missing values with the mean of all recorded values would not be an accurate representation of the data. To combat this, 'ACT Score' was turned into an indicator variable which only reports whether or not a student reported an ACT score, not what the numerical value of the score was.

Students with a cumulative GPA that was either missing or equal to 0 were removed from the data set. Most students who had a 1st year GPA of 0.00 also had missing values for 2nd, 3rd, and 4th year GPAs, so those students were removed from the study. If students had missing values for the GPA of every year, they were also removed. Many students for various reasons transfer into and out of Worcester State University. This, unfortunately, creates many missing values, however, unlike HS GPA and SAT scores, these missing values cannot be replaced with the mean of every reported value. Suppose, for example, a student had a GPA of 0.20 in their first year and subsequently dropped out or transferred. Replacing the 2nd, 3rd, and 4th year GPAs with the mean of each year (approximately 3.10) would not produce an accurate view of this student's grades. By replacing the missing GPA of each year with the student's *cumulative* GPA, however, a more exact view of this student's grades is produced.

After fixing the columns with missing values, there are still some variables that need to be altered before the data can be visualized and analyzed. Every variable must be numeric for linear regression to work, so many variables needed to be turned into indicator variables. Majors, for example had to be turned into a series of indicator variables. Each major became a new variable that had a value of either 0 or 1 - the variable was 1 when the student had this major. For example, a math major would have a 1 in the Math column and a 0 value for every other major. Similarly, gender, race/ethnicity, first generation, housing, state, and county needed to be turned into indicator (0 or 1) variables. Thus, 29 variables became 105 variables.

The remaining classifiers besides linear regression are used to predict a certain category instead of a numerical value. Because of this, we classified cumulative GPA into the following ranges: low (0 - 1.99], mid (1.99 - 3.29], and high (3.29 - 4.00], as seen in Table 1. Rule-based classifiers, decision tree classifiers, and other data mining techniques can then predict which cumulative GPA class a student falls into.

Modeling

Data mining techniques such as linear regression, PART, OneR, ZeroR, J48, REPTree, RandomTree, Random Forest, K-Nearest Neighbors, and Naïve Bayes were used to determine which factors contribute to student success. Each of these algorithms output a model and the accuracy of each model is weighed against each other to find the best model for our data set. For all modeling techniques, the data set was split into two groups; a random 80% of the data was used to build the model and the remaining 20% was used to test the model for accuracy once it was built.

Linear Regression

Linear regression creates an equation using given factors that models the desired variable. To find the most accurate linear model, we can 'step through' each combination of the given variables to find what factors with what coefficients, or weights, most accurately predict the value of the desired variable. For example, to form the best predictor model for cumulative GPA based on 1st, 2nd, 3rd, and 4th year GPAs, our code will examine each combination of those variables (1st year GPA; 1st and 2nd year GPA; 1st and 3rd year GPA; ...; 1st, 2nd, 3rd, and 4th year GPA) to find the most accurate model.

Many factors contribute to the accuracy of a linear model. Performing linear regression on a data set of a large enough size makes it possible to predict the numerical value for cumulative GPA with high accuracy. Once we step through each combination of the variables we input and find the best model, we must examine the p-values, Multiple R-Squared, Adjusted R-Squared, Min-Max Accuracy, and Mean Absolute Percentage Error (MAPE) of the model.

P-Value: To fully comprehend p-value, we must first understand null and alternative hypotheses. A linear model assumes the null hypothesis - that the correlation between the given variables and the desired variable is zero. In the case of our study, this means that linear regression assumes that the factors do not impact cumulative GPA at all. The p-value represents the probability that, given that the null hypothesis is true, the observed results occurred. In other words, the p-value tells us how likely it would be for our data set to occur by chance with no correlation between variables. As we are aiming to determine whether there is a correlation between our given variables and a student's cumulative GPA, a p-value of less than 0.05 will show us that our model is statistically significant, showing that it is less than 5% likely that our data set occured by chance; our variables factor into the value of cumulative GPA. The smaller the p-value, the more significant our results are.

R-Squared: The R-Squared value shows us what proportion of the total variability of our desired variable is explained by our linear model. In other words, the R-Squared value tells us how well our model explains the data in our data set. The higher the R-Squared value, the higher the proportion of variability that our model predicts.

Adjusted R-Squared: Adjusted R-Squared is the same as R-Squared, however it accounts for the number of variables used in the model. A more complex model will never be less accurate than its subset, but a complex model is not necessarily better for analysis in the long run. Data analysts prefer models that have a balance between high accuracy and simplicity, thus R-Squared is discounted based on the number of variables to produce an Adjusted R-Squared value.

Min-Max Accuracy: After creating a model to examine, we can then compare the values that the model predicts to the observed data. To create our model we used 80% of the data to create a 'training set' and we used the remaining 20% to test the model. By comparing the predicted cumulative GPAs with the actual cumulative GPAs provided by the remaining 20% of our data set, we can produce a Min-Max Accuracy for the model. This value is found by comparing the actual and predicted values for each row of the data set and dividing the minimum by the maximum. We then find the average of these values across each row. Simply, the Min-Max Accuracy tells us how far off the prediction is from the actual values overall. A higher Min-Max Accuracy value means a higher accuracy rate for the model.

MAPE: The Mean Absolute Percentage Error, like the Min-Max Accuracy, measures the precision of the model based on predicted vs. actual values. In contrast to Min-Max Accuracy, however, MAPE is a measure of the average absolute percentage error of each row of the data set. By finding the difference between predicted and actual value, finding the absolute value of that error as a percent, then finding the mean error across the whole data set, we can find the MAPE. Because this value is a representation of error, a smaller MAPE shows that the model is better than one with a higher MAPE and therefore higher error. The closer to 0% the Mean Absolute Percentage Error is, the more accurate the model.

Rule-Based Models: PART, OneR, and ZeroR

Rule-based classifiers create rules based on a specified percentage of the data set. For example, an 80% split uses 80% of the data to build a model and uses the remaining 20% to test that model. These models are composed of a series of if-else statements that determine how a data point will be categorized. For example, in our study, students will be classified into one of the following cumulative GPA classes: low, mid, or high, as seen in Table 1, using given factors. The rules developed by the classifier will determine which class the student will fall into. A very simple example is this: IF a student has a 4th year GPA of greater than 3.35, THEN the student's cumulative GPA will belong in class 'high'. If the student does *not* have a 4th year GPA that is greater than 3.35, then the classifier will move on the next rule it develops.

Class	Cumulative GPA Range
Low	[0, 1.99]
Mid	(1.99, 3.29]
High	(3.29, 4.0]

Table 1: Cumulative GPA Ranges by Class

PART in particular as a rule-based classifier takes every variable and takes them all into account when determining which rules produce the most accurate model. Because of this, there can be many complicated PART rules which create a model with high accuracy that may be more complicated.

OneR, short for 'One Rule', generates one rule for each predictive variable and then chooses the rule with the least error as its overarching rule. This can create extremely simple yet accurate models to use to classify instances into groups.

ZeroR, although technically a rule-based classifier, ignores all predictive variables and simply creates a single rule that predicts an outcome based solely on the majority class. For example, if 99% of our data set belonged in the 'high' class, ZeroR would predict that every new instance would also belong in the 'high' class with 99% accuracy. However, since ZeroR does not examine any correlation or statistical significance, this precision rate is not a good indicator for whether or not ZeroR is a sensible model for this study. If there are more instances of GPAs in the 'high' class than any other class, ZeroR will always classify new instances as belonging in the 'high' group. Because of this, ZeroR is often far less accurate than other classifiers and any accuracy is based solely on distribution of values, not on predicting contributing factors.

Decision Tree Algorithms: J48, REPTree, and Random Forest

A decision tree is, as its name suggests, a tree-like model of decisions and their outcomes. For example, if, in a simple hypothetical model with one decision, a student had a 4th year GPA of less than or equal to 3.25, then they were likely to get a cumulative GPA in the 'mid' range (1.99 - 3.29]. The decision tree would have two branches, one for if a student got a 4th year GPA of less than or equal to 3.25 that results in the student having a cumulative GPA in the 'mid' range, and one where the student has a 4th year GPA of greater than 3.25 resulting in having a GPA in the 'high' range. The decision tree will also include the rates at which each outcome will happen. Using decision trees, one study cited 94.40 percent accuracy.

One type of decision tree algorithm, J48, also known as the C4.5 algorithm, chooses the predictor variable that produces the greatest gain of information, uses that variable as the node, then splits that variable onto the variable with the next greatest information gain as the start of the branches. More in-depth and detailed information about J48 and decision trees can be found in [3].

REPTree generates a decision tree, but then 'prunes' the tree by examining each subtree and seeing if it can be replaced by a single node without significantly lowering the accuracy of the tree. This model, despite its accuracy, can be seen as aggressive as it can possibly take out nodes that are significant.

Random Forest, another type of decision tree classifier, is made up of a number of individual decision trees that each output a predicted outcome. The outcome that has been produced the greatest number of times by the individual trees is the final predicted output.

K-Nearest Neighbors

K-Nearest Neighbor is an algorithm which randomly generates a number of nodes, which each instance is then compared to. Each instance is grouped with whichever node it is 'closer', or most similar, to. The most frequent class out of the grouped instances is then the label of that node, and any new instances will be classified as that class when grouped with the node in question. More information about kNN can be found in [8].

Naïve Bayes

The Naïve Bayes classifier is the only probabilistic classifier used in this study. This classifier assumes that factors are unrelated to one another, which is not accurate. SAT Total, one of our variables, is very closely related to SAT Math and SAT Verbal. Naïve Bayes assumes that there is zero correlation between these variables, which is a naïve assumption. It uses Bayesian statistics to find the likelihood that an instance belongs in each class using conditional probability and Bayes' Rule. Bayes' Rule, in Equation 1, whose derivative is outside of the scope of this project, uses probability rules that produce the conditional probability that event A occurs given that event B occurs. Naïve Bayes classifier uses Bayes Theorem to predict the probability that an instance is in class y_i given that predictive factors $x_1, x_2, ..., x_n$ occurred. Y_i replaces A in Bayes Rule and $x_1, x_2, ..., x_n$ replaces B, as seen in Equation 2. The probability for each class is calculated, then the class with the highest likelihood is assigned to that instance. The main drawback for using the Naïve Bayes classifier is its assumption that $x_1, x_2, ..., x_n$ are all independent factors, which is not usually true in practice. The derivation of Bayes Rule and more information about Naïve Bayes and Bayesian statistics can be found in [9].

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)} \tag{1}$$

$$P(y_i|x_1, x_2, ..., x_n) = \frac{P(x_1, x_2, ..., x_n|y_i) \cdot P(y_i)}{P(x_1, x_2, ..., x_n)} \tag{2}$$

Results

Linear Regression

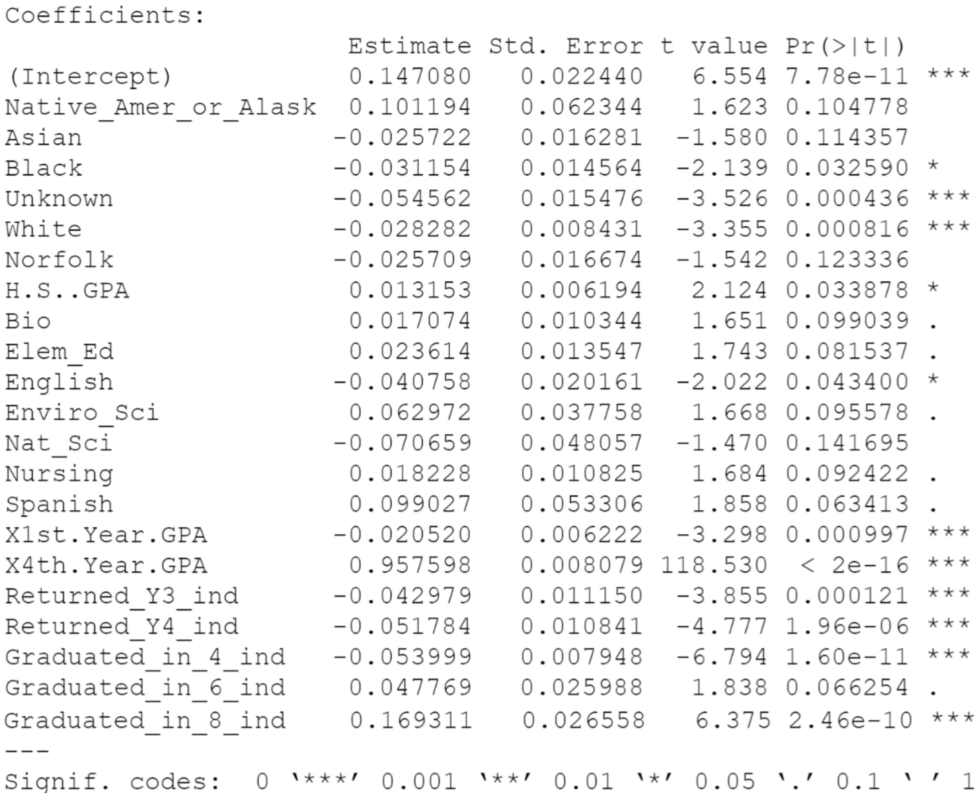
The data was split into a training model and test model using a randomly selected 80% of the data to build the model and the remaining 20% to test the model. Using 80%, we built a linear regression model, Figure 4, using R, which takes into account every variable given and then develops a model using the subset of those variables that describes the data with the highest accuracy. The coefficients in the 'Estimate' column in 5 are multiplied by each variable in the model in Figure 4 to predict the students' final GPAs. The variables that are the most significant contributions to cumulative GPA are indicated by asterisks: the more asterisks, the more significant. Coefficients such as 1st Year GPA, 4th Year GPA, Returned Year 3, Returned Year 4, Graduated in 4, and Graduated in 8 are the most significant academic predictors in the model. Pre-admission and demographic features that most significantly contribute to cumulative GPA are whether a student has an Unknown or White race or ethnicity. Variables with a value of less than 0.05 in the p-value, or $Pr(> |t|)$, column are not considered statistically significant.

Cum.GPA = Native.Amer.or.Alask + Asian + Black + Unknown + White + Norfolk + H.S..GPA + Bio + Elem.Ed + English + Enviro.Sci + Nat.Sci + Nursing + Spanish + 1st.Year.GPA + 4th.Year.GPA + Returned.Y3 + Returned.Y4 + Graduated.in.4 + Graduated.in.6 + Graduated.in.8

Figure 4: Best linear regression using 80% of data set in R; remaining 20% of the data used to test this model

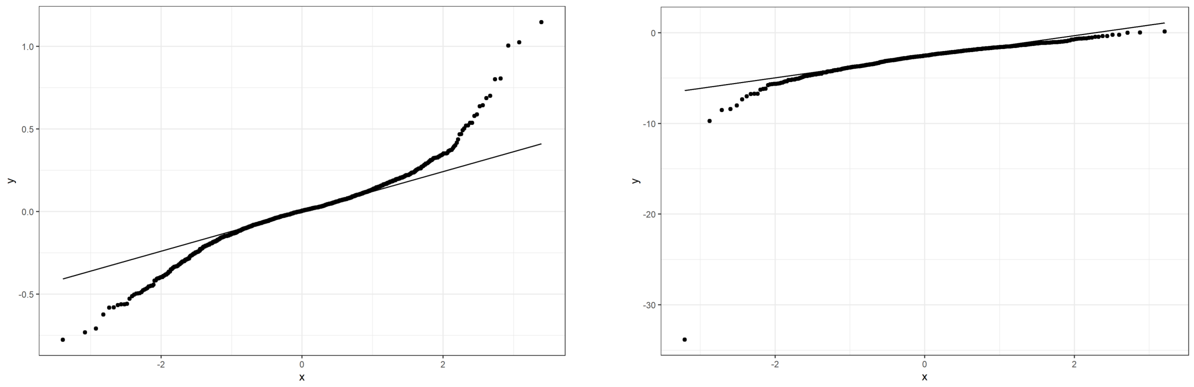
Of the 7 majors included in the model, 4 are STEM majors and, of those STEM majors, 3 *increase* the cumulative GPA by some degree (the coefficients in Figure 5 associated with these variables are positive).

For this linear regression model, the p-value is less than $2.2e-16$; the p-value is essentially zero, so the results are statistically significant - at more than a 99% significance, the data does not follow this model solely by chance. The R-Squared and Adjusted R-Squared values are 0.9705 and 0.9701 respectively, meaning more than 97% of the variation in the data is explained by the model. Adjusted R-Squared is not significantly lower than the R-Squared value, meaning the model is not extremely complicated and is therefore desirable.



```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           0.147080   0.022440   6.554 7.78e-11 ***
Native_Amer_or_Alask  0.101194   0.062344   1.623 0.104778    
Asian                -0.025722   0.016281  -1.580 0.114357    
Black                -0.031154   0.014564  -2.139 0.032590 *  
Unknown              -0.054562   0.015476  -3.526 0.000436 ***
White                -0.028282   0.008431  -3.355 0.000816 ***
Norfolk              -0.025709   0.016674  -1.542 0.123336    
H.S..GPA              0.013153   0.006194   2.124 0.033878 *  
Bio                   0.017074   0.010344   1.651 0.099039 .  
Elem_Ed               0.023614   0.013547   1.743 0.081537 .  
English              -0.040758   0.020161  -2.022 0.043400 *  
Enviro_Sci            0.062972   0.037758   1.668 0.095578 .  
Nat_Sci              -0.070659   0.048057  -1.470 0.141695    
Nursing               0.018228   0.010825   1.684 0.092422 .  
Spanish               0.099027   0.053306   1.858 0.063413 .  
X1st.Year.GPA        -0.020520   0.006222  -3.298 0.000997 ***
X4th.Year.GPA         0.957598   0.008079 118.530  < 2e-16 ***
Returned_Y3_ind      -0.042979   0.011150  -3.855 0.000121 ***
Returned_Y4_ind      -0.051784   0.010841  -4.777 1.96e-06 ***
Graduated_in_4_ind   -0.053999   0.007948  -6.794 1.60e-11 ***
Graduated_in_6_ind    0.047769   0.025988   1.838 0.066254 .  
Graduated_in_8_ind    0.169311   0.026558   6.375 2.46e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5: Linear regression coefficients and significance of model using 80% of data set



(a) Original QQ-plot of residuals

(b) QQ-plot of log of residuals

Figure 6: Original and transformed QQ-plots of residuals of best linear regression model

Using the remaining 20% of the data, we evaluated the precision of the model. The Min-Max Accuracy is 0.9784, so the model is highly accurate at 97.84%. The MAPE, or Mean Absolute Percentage Error, is 2.22%, meaning the average error for the model in each row is 2.22%, which is very low. This means that any prediction has an error of about 2.22%.

Based on these factors, our model is highly accurate. We must, however, also examine at the normality of the residuals of the model. It is not enough to say that, for example, our MAPE is low and thus the model is accurate. Checking for normalcy in our residual values tells us the validity of our model's predictions. When running a linear model, the residuals are assumed to be normally distributed, however, we must verify this assumption. Unlike the MAPE, the absolute value is not taken to determine normalcy in residuals. To be considered normal, residuals must form roughly a straight line. Figure 6a shows that our residuals in their original form are not normal, since they do not lie on the line generated in the plot, they have heavy tails. By transforming the residuals by, for example, taking their log and plotting them, we can make them slightly more linear, however, the log-residuals are still left-skewed enough to be considered not normal. However, normalcy is more or less subjective, so determining whether residuals are definitively normal is difficult in most cases.

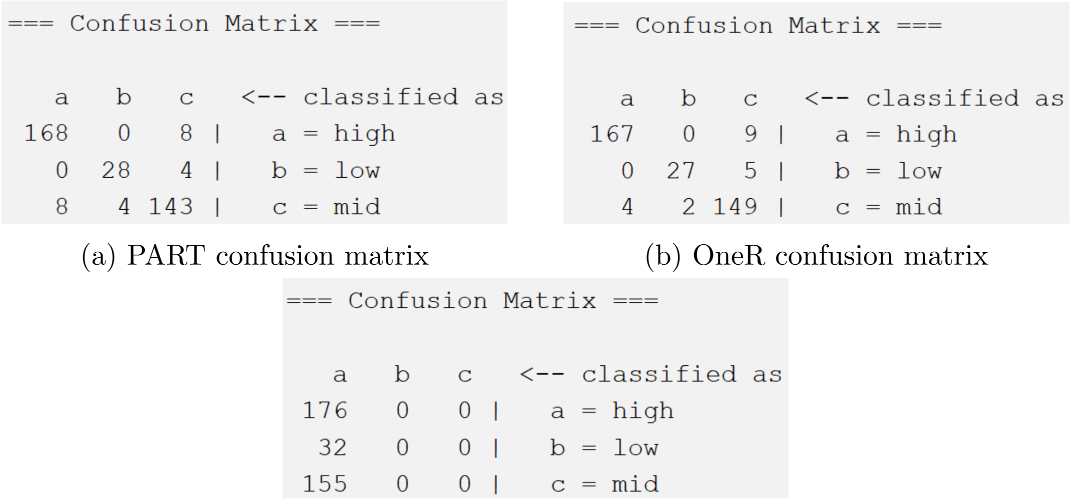
When residuals are not normally distributed, this means that linear regression may not be the best model type for this data. Although our model is highly accurate using our 80/20 split of data, it becomes less accurate when predicting data that is extremely low; our model is also not as good at determining when a student has an extremely high GPA, but there are more students on the higher end of the spectrum than the lower end. On the low end, there are not enough of this type of student in our data set to provide accurate predictions of students in this extreme.

In Figure 1, we can see that the histogram of cumulative GPA is left-skewed and therefore not normally distributed, and the QQ-plots in Figure 6 supports that abnormality. Because we have far less information for students with GPAs in the 'low' range, our prediction model will not work as well to output students with lower GPAs. The model is, however, still extremely accurate, it is just less accurate in predicting GPAs in the ‘low’ range, which is understandable due to the fact that students are not accepted to state universities such as this one with a high school GPA lower than 2.0, which may eliminate many students who are prone to low GPAs from the data set entirely.

Ruled-Based Classifiers

Rule-based classifiers such as PART, OneR, and ZeroR were used to predict which GPA category (low [0 - 1.99], mid (1.99 - 3.29], and high (3.29 - 4.00]; Table 1) each student belongs in based on predictive variables. PART, as explained above, makes use of every variable and develops a rule list containing the series of rules that best describe the data. When running PART on this data set, a series of 33 rules were developed to best predict GPA class. These rules, after using 80% of the data to build the model and 20% to test it, correctly classified 93.11% of test instances. The confusion matrix for PART in Figure 7a shows how many instances were correctly classified into each group. The values on the diagonal show the correctly classified instances using the rules developed by the PART classifier. Any values *not* on the diagonal show instances that were incorrectly classified and which groups they were sorted into instead.

Another rule-based classifier, OneR, generates one rule for each variable and determines which rule has the highest accuracy, then uses that as the predictor rule for the model. The best rule determined by OneR is the following: If the 4th Year GPA is < 1.995, then class: ‘low’; if 4th Year GPA is < 3.275, then class: ‘mid’; if neither of those are true, 4th Year GPA is ≥ 3.275, then class: ‘high’. Using this rule, OneR has 94.4% of its instances classified correctly, which, although this model is far simpler than PART, has a higher accuracy rate than the PART decision list of 33 rules. The confusion matrix in Figure 7b shows how many instances are classified correctly into which classes.



```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 168   0   8 |   a = high
   0  28   4 |   b = low
   8   4 143 |   c = mid
```

(a) PART confusion matrix

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 167   0   9 |   a = high
   0  27   5 |   b = low
   4   2 149 |   c = mid
```

(b) OneR confusion matrix

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 176   0   0 |   a = high
  32   0   0 |   b = low
 155   0   0 |   c = mid
```

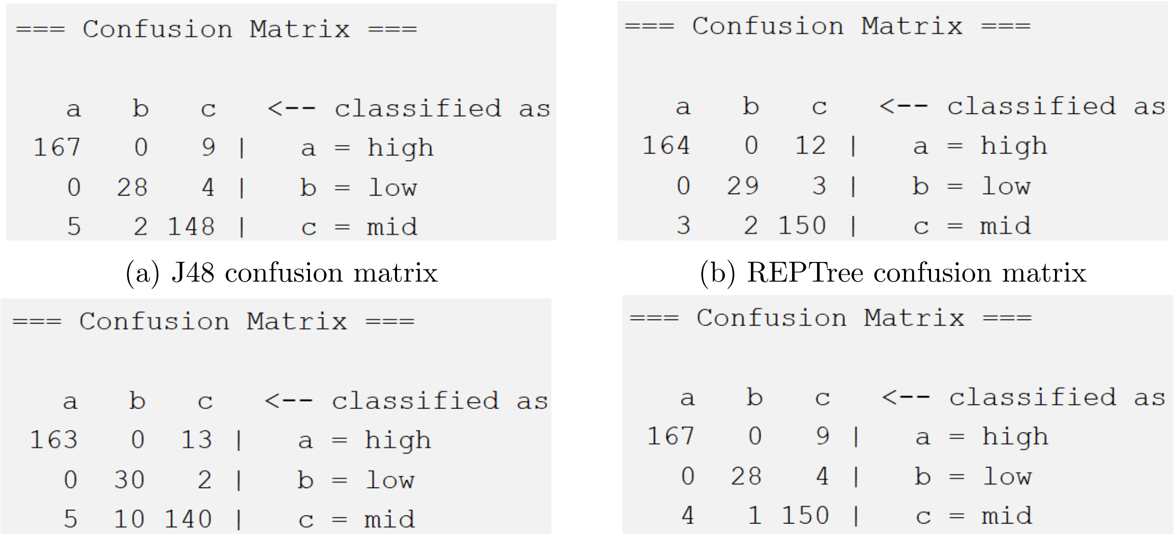
(c) ZeroR confusion matrix

Figure 7: Confusion matrices for rule-based classifiers

ZeroR generates a pseudo-rule that chooses the class with the majority of instances and assigns every new instance to that class. ZeroR determined that the randomly split 80% of the data set was composed mostly of GPAs in the 'high' class, so every instance in the test data set was assigned to class 'high'. This only had a 48.48% accuracy, since only 48.48% of the test data set was actually in class 'high'. This can be seen clearly in Figure 7c. ZeroR only looks at the distribution of the data and doesn't take into account any correlation or statistical significance, which is not desirable for this study.

Decision Trees

Decision trees algorithms such as J48, REPTree, and Random Forest generate models of a tree-like structure which contain 'branches' which each new instance would follow until they are classified into a group. Each branch is made up of nodes which break off into more branches until a conclusion is reached for each branch. Every possible situation must be accounted for in each decision tree, so they can become complicated as they get bigger and more nodes. Simplicity and accuracy must be balanced when deciding which decision tree is the best for data modeling.



```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 167   0   9 |   a = high
   0  28   4 |   b = low
   5   2 148 |   c = mid
```

(a) J48 confusion matrix

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 164   0  12 |   a = high
   0  29   3 |   b = low
   3   2 150 |   c = mid
```

(b) REPTree confusion matrix

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 163   0  13 |   a = high
   0  30   2 |   b = low
   5  10 140 |   c = mid
```

(c) Random Tree confusion matrix

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 167   0   9 |   a = high
   0  28   4 |   b = low
   4   1 150 |   c = mid
```

(d) Random Forest confusion matrix

Figure 8: Confusion matrices for decision tree algorithms

The J48 classification tool, also known as the C4.5 algorithm, produced an accuracy rate of 94.49%. Figure 9 shows us the decision tree that the J48 classifier produced, which is somewhat complicated, but not overwhelmingly so. Its confusion matrix in Figure 8a displays the number of correctly and incorrectly classified instances for each class of cumulative GPA using 20% of the original data set to test the classification model.

The REPTree classifier also produced an accuracy rate of 94.49%, the same as the J48 classifier's accuracy rate. As seen in Figure 10, however, the decision tree output is different from that of J48. When two models have the same accuracy rate, one must choose the simpler model, since simpler is better for data modeling and analysis; why use a complicated model when a simple one is equally as accurate? Because the simpler model is better, REPTree is a more desirable model than J48 is in this case despite the fact that they have the same accuracy rate.

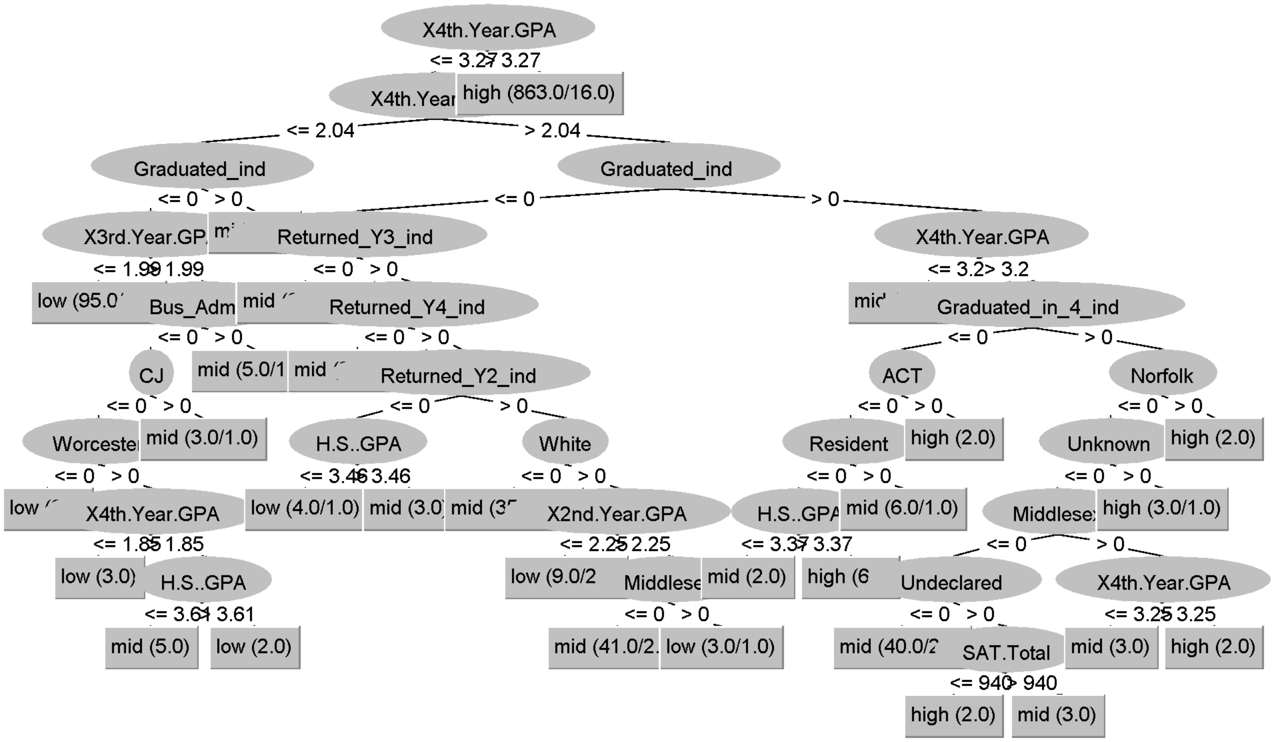


Figure 9: J48 decision tree output

Random Tree, another decision tree algorithm, is one component of the Random Forest algorithm. Random Forest uses many randomly generated decision trees, Random Trees, and chooses the most-generated class as the prediction for new instances. The Random Tree algorithm on its own has a lower accuracy than the Random Forest algorithm since Random Tree is a subset of Random Forest. Random Tree on its own is not one of the better classifiers, as seen in Figure 11, because it is far more complicated than any of the other decision tree models, but only has a accuracy rate of 83.75%.

Random Forest, on the other hand, is made up of many Random Tree decision trees and the most frequent outcome among many of them is deemed the final classification of the instance. Because Random Forest is made up of a random number of randomly generated decision trees, it cannot be easily visualized. Comparing the confusion matrices for Random Tree and Random Forest, Figure 8c and Figure 8d, respectively, we can see that Random Forest does a better job of accurately predicting the GPA class of test instances. Random Forest has an accuracy rate of 93.66% compared to Random Tree's 83.75%.

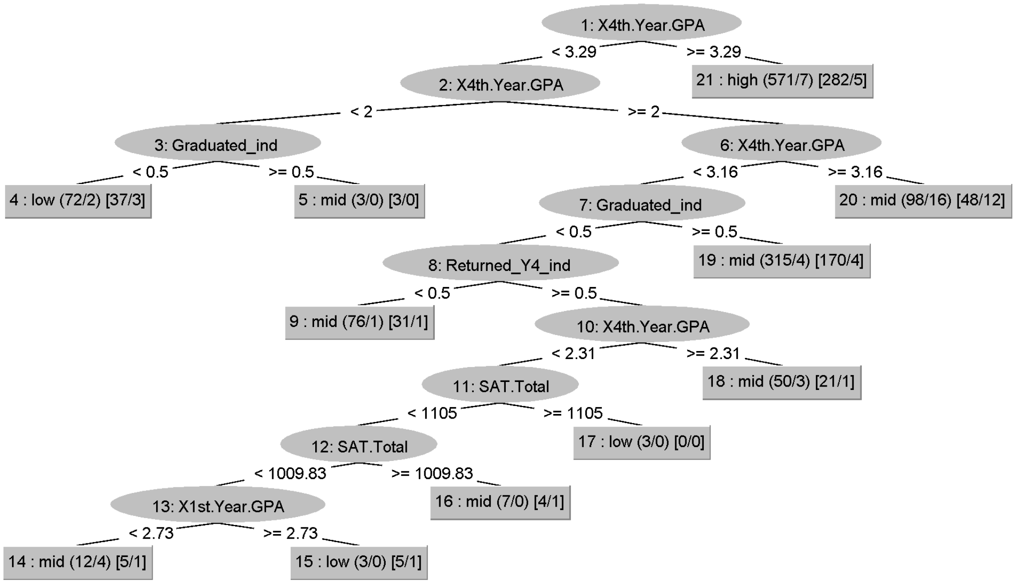


Figure 10: REPTree decision tree output

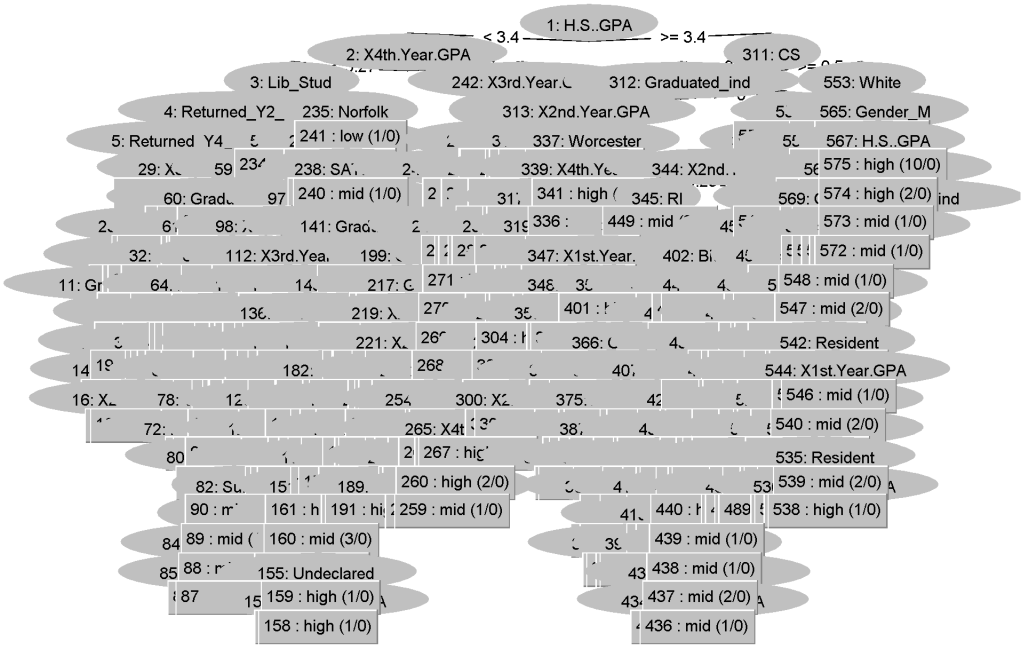
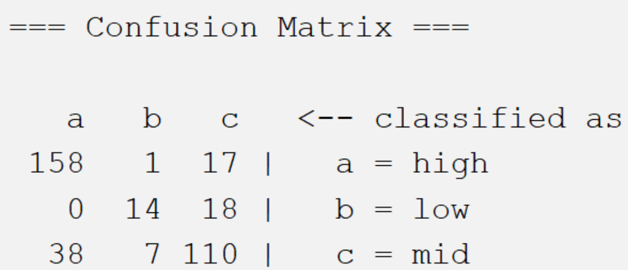


Figure 11: Random Tree output

K-Nearest Neighbors

When running the K-Nearest Neighbor algorithm, we must determine which k value maximizes the algorithms accuracy. By cycling through k values, we found that, when $k = 10$, the algorithm produced its highest accuracy of 77.686%. This is the lowest accuracy rate out of all of the data mining tools so far, and the confusion matrix can be seen in Figure 12. This clustering technique, in this case, was not as accurate as the other techniques used, and thus is not the best model for predicting student success and its contributing factors.



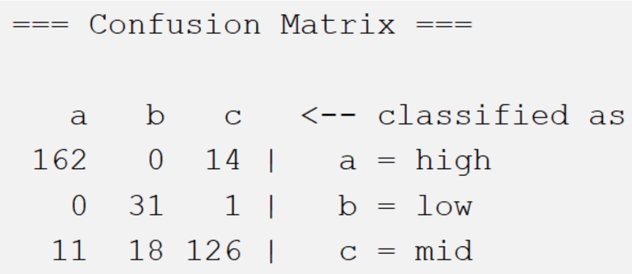
```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 158   1  17 |   a = high
   0  14  18 |   b = low
  38   7 110 |   c = mid
```

Figure 12: Confusion Matrix of K-Nearest Neighbors, $k = 10$

Naïve Bayes

The Naïve Bayes classifier uses Bayesian statistics, specifically Bayes' Rule, to determine the probability that an instance falls into a certain class given that each factor has a certain value. Referencing back to Equation 2 shows us how Naïve Bayes uses Bayes' Rule (Equation 1) to determine which class each new instance falls into. Evaluating our model using our 20% test data set, the Naïve Bayes classifier produces an accuracy rate of 87.8788%. The confusion matrix for Naïve Bayes can be found in Figure 13.



```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 162   0  14 |   a = high
   0  31   1 |   b = low
  11  18 126 |   c = mid
```

Figure 13: Naïve Bayes confusion matrix

Accuracy of Rule-Based Classifiers	
Classifier Name	Accuracy Rate (%)
PART	93.3884
OneR	94.4904
ZeroR	48.4848

Table 2: Rule-Based Classifier Accuracy Rates

Accuracy of Decision Tree Classifiers	
Classifier Name	Accuracy Rate (%)
J48	94.4904
REPTree	94.4904
Random Tree	91.7355
Random Forest	95.0413

Table 3: Decision Tree Algorithm Accuracy Rates

Accuracy of Linear Regression			
Classifier Name	R-Squared (%)	Adjusted R-Squared (%)	Min-Max Accuracy (%)
Linear Regression	97.05	97.01	97.8439

Table 4: Linear Regression Accuracy Rates

Accuracy of Probabilistic Classifier	
Classifier Name	Accuracy Rate(%)
Naïve Bayes	87.8788

Table 5: Naïve Bayes Accuracy Rate

Classifier Name	Accuracy Rate(%)
K-Nearest Neighbors, $k = 10$	77.686

Table 6: K-Nearest Neighbor Accuracy Rate, $k = 10$

Discussion

These results answer our first research question:

Can we build accurate models to predict students' cumulative GPA both numerically and by category?

Our linear regression model predicted numerical values of cumulative GPA based on academic and demographic factors with an average absolute error of 2.22%. The ruled-based classifiers, decision trees, probabilistic classifier, and K-Nearest Neighbors classifier predicted the class of cumulative GPA into which a student will fall given a series of predictive factors. These classes are seen in Table 1 and the accuracy of these models are clearly seen in Tables 2, 3, 4, 5, and 6. Linear regression has the highest accuracy rate overall at 97.84%, and Random Forest had the second-best model with a 95.04% accuracy rate.

Our final, most accurate model, produced using linear regression, answers our second research question, as follows:

What factors most accurately predict students' final cumulative GPAs?

The most accurate linear regression model (Figure 4) exhibits the factors that most contribute to student cumulative GPA. The final linear regression model, complete with coefficients, can be seen in Figure 14. We can see by this figure and the presence of asterisks in Figure 5 that the most statistically significant predictive factors that contribute to cumulative GPA are White, race and/or ethnicity Unknown, 1st year GPA, 4th year GPA, Returned Y3, Returned Y4, Graduated in 4, and Graduated in 8. The remaining factors do contribute to this prediction, however these factors do not have p-values that are significantly lower than 0.05. Factors like Black, H.S. GPA, and English are also predictors with p-values in the 95% significance range, however they are still less predictive than variables with lower p-values. This model overall has the highest possible accuracy, however each individual factor may not significantly contribute to the model.

$$\begin{gathered}\mathbf{Cum.GPA = 0.147080 + (0.101194)*Native.Amer.or.Alask +} \\ \mathbf{(-0.025722)*Asian + (-0.031154)*Black + (-0.054562)*Unknown +} \\ \mathbf{(-0.028282)*White + (-0.025709)*Norfolk + (0.013153)*H.S..GPA +} \\ \mathbf{(0.017074)*Bio + (0.023614)*Elem.Ed + (-0.040758)*English +} \\ \mathbf{(0.062972)*Enviro.Sci + (-0.070659)*Nat.Sci + (0.018228)*Nursing +} \\ \mathbf{(0.099027)*Spanish + (-0.020520)*1st.Year.GPA + (0.957598)*4th.Year.GPA} \\ \mathbf{+ (-0.042979)*Returned.Y3 + (-0.051784)*Returned.Y4 +} \\ \mathbf{(-0053999)*Graduated.in.4 + (0.047769)*Graduated.in.6 +} \\ \mathbf{(0.169311)*Graduated.in.8}\end{gathered}$$

Figure 14: Linear regression model with coefficients

English is one of two majors (English and Natural Sciences) in this model that decrease the cumulative GPA if enrolled in this major. This can occur for multiple reasons. Either it is by the nature of the students who generally enroll in these majors, the difficulty of the course material, or the stringency of grading throughout the major.

Four out of the seven majors that contribute to this model are STEM majors. As Tomkin and West (2022) describe in their study, there are many STEM majors which choose that academic pathway *because* of their talent and scholastic prowess. The nature of the students in these majors could explain why, despite the rigor of STEM majors, having these majors indicates an increase in cumulative GPA.

It is clear why H.S. GPA would contribute to a student's cumulative GPA at the end of their time in college. Students whose GPA in high school was very high may have a higher-than-average GPA in college. However, students do not always stay constant in their academic career, which explains why the correlation between H.S. GPA and cumulative GPA is not as high as it is for other variables.

All race/ethnicity factors included in the model (Asian, Black, Unknown, White) except one (Native American or Alaska Native) decrease the cumulative GPA model. One explanation for why the Native American or Alaska Native variable predicts a higher GPA is that, because there are so few Native American students in our data – only 4 –, these students may have had very high GPAs that were potentially influenced by other factors as well. Because our model was built only using existing data, the only information available to construct this model included these Native American students with high GPAs, thus our model assumed there was a correlation between the Native American or Alaska Native race/ethnicity and a higher GPA.

A student's 4th year GPA increases the final cumulative GPA by the highest factor out of all of the predictive variables. This and the significance of the 1st year GPA may be explained by the fact that students in their 1st year are adjusting to their new environment and to an academically demanding environment that they may not be used to. Because of this and other potential reasons, 1st year GPA can bring down a student's cumulative GPA. A student's 4th year GPA may have a higher weight in predicting cumulative GPA because, by the end of their time in college they have a more consistent GPA which is a more accurate representation of their skills and academic habits as a student.

This information can be extremely useful to administration, as well as to students and faculty. Although current students enrolled in Worcester State University can not necessarily get an accurate prediction of their own cumulative GPA, administrators can examine these predictive variables and their underlying causes and find ways to combat factors that cause lower GPAs and provide support to students who interact with these factors.

Conclusion

Using linear regression, PART, OneR, ZeroR, J48, REPTree, Random Tree, Random Forest, K-Nearest Neighbors, and Naïve Bayes algortihms, we developed a highly accurate model used to predict cumulative GPA from students' academic and demographic pre-admissions and collegiate factors. We were able to determine that White, race/ethnicity Unknown, 1st year GPA, 4th year GPA, Returned Y3, Returned Y4, Graduated in 4, and Graduated in 8 predictive variables contribute the most to approximating cumulative GPA at Worcester State University. This information can be used to help faculty and administration provide support in areas that lower GPA and continue working towards helping students in areas that increase GPA. For example, by supporting the Retention Office and understanding why students may leave the university early, more students may graduate from Worcester State University within 6 or 8 years (graduate school or continuing education), increasing the success of those students. This information can be used to provide support to students to increase their GPAs, consequently raising the average GPA of the school and improving the school in general.

References

[1] Bilal, Muhammad, et al. *The Role of Demographic and Academic Features in a Student Performance Prediction.* Scientific Reports, 2022.

[2] Eibe Frank, Mark A. Hall, and Ian H. Witten. *The WEKA Workbench, Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann. Fourth Edition, 2016.

[3] Khanna, Nilima. *J48 Classification (C4.5 Algorithm) in a Nutshell.* Medium.com, 2022. https://medium.com/@nilimakhanna1/j48-classification-c4-5-algorithm-in-a-nutshell-24c50d20658e.

[4] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, 2022. https://www.R-project.org/.

[5] Stadlman, Margaret, et al. *Artificial Intelligence Based Model for Prediction of Students' Performance: A Case Study of Synchronous Online Courses during the COVID-19 Pandemic.* Journal of STEM Education: Innovations and Research. Volume 23, no. 2, 2022.

[6] Tomkin, Jonathan H., and Matthew West. *STEM Courses are Harder: Evaluating Inter-Course Grading Disparities with a Calibrated GPA Model.* International Journal of STEM Education. Volume 12, no. 1, 2022.

[7] Yağcı, Mustafa. *Educational Data Mining: Prediction of Students' Academic Performance using Machine Learning Algorithms.* Smart Learning Environments. Volume 9, no. 1, 2022.

[8] Yıldırım, Soner. *K-Nearest Neighbors (kNN) — Explained.* Towards Data Science, 2020. https://tinyurl.com/4r5xw2jt.

[9] Yıldırım, Soner. *Naive Bayes Classifier Explained.* Towards Data Science, 2020. https://towardsdatascience.com/naive-bayes-classifier-explained-50f9723571ed.