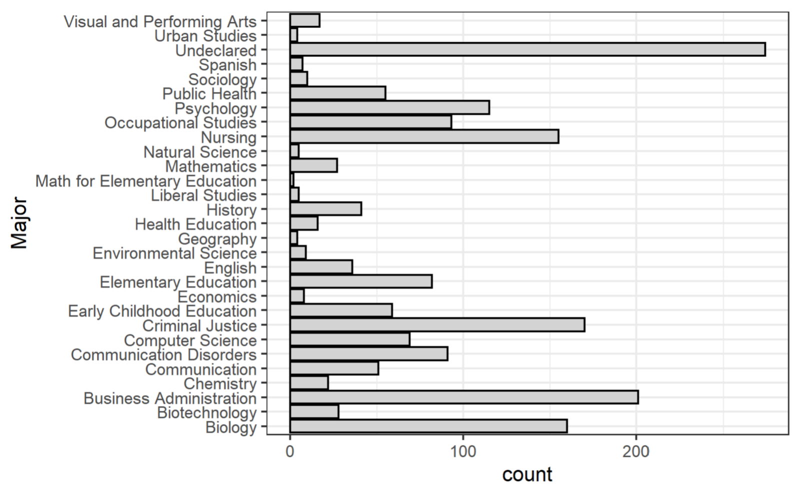
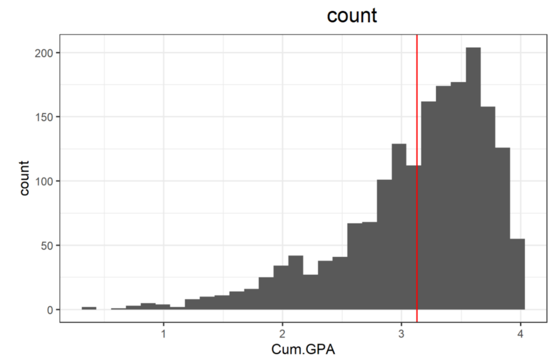
Predicting Student Performance Using Data Mining Techniques

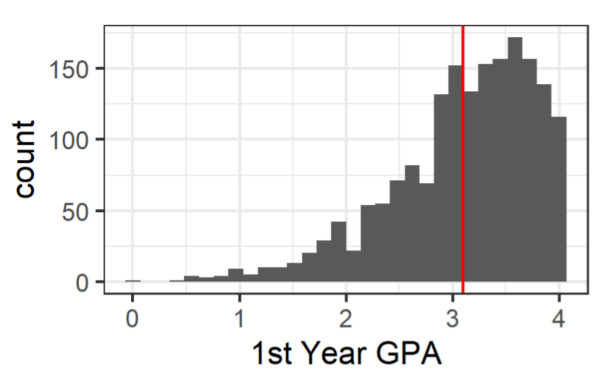
Avery Oldakowski | Jason Hardin – Project Advisor
aoldakowski@worcester.edu | Worcester State University

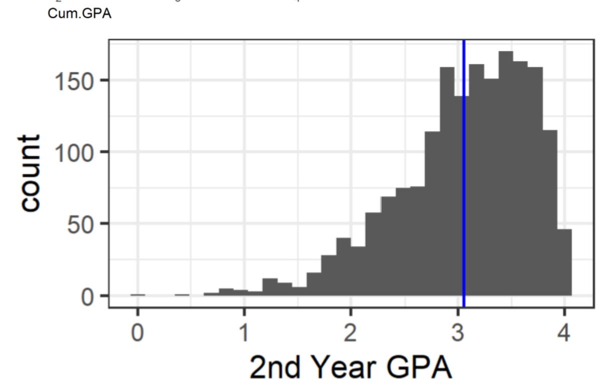
Participant Academic Information

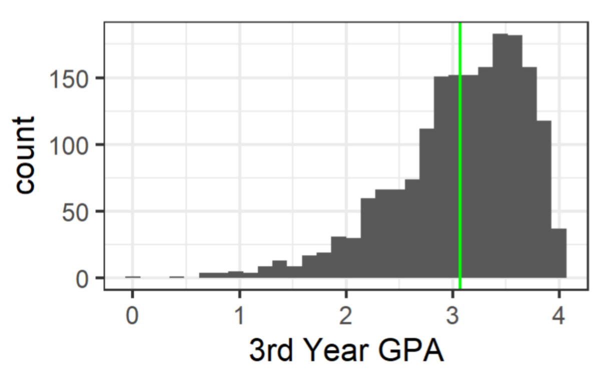
Data was collected from 1816 students from Worcester State University whose start years are 2014, 2015, and 2016. Information from their entire time at the university was recorded and used in this study.

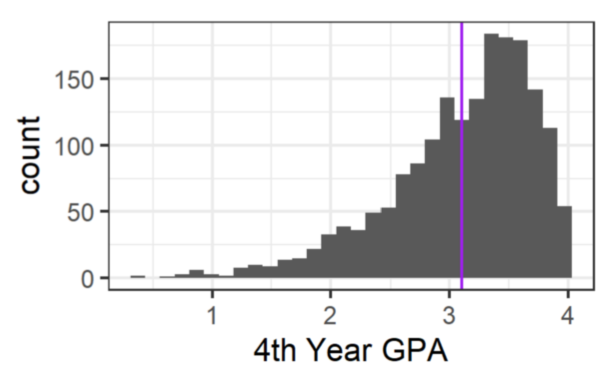




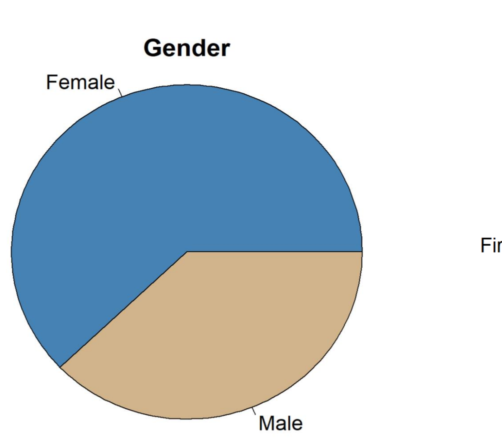


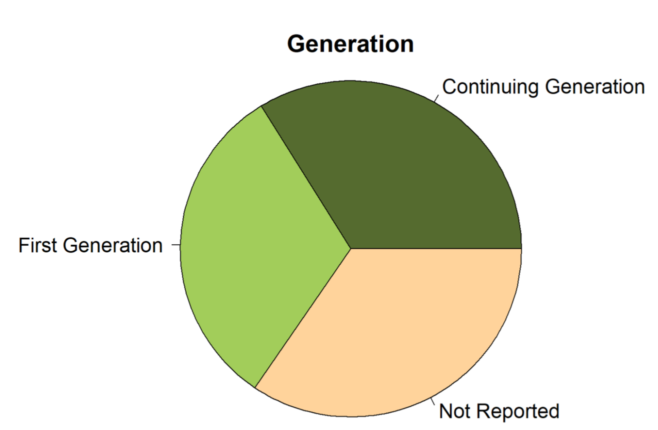


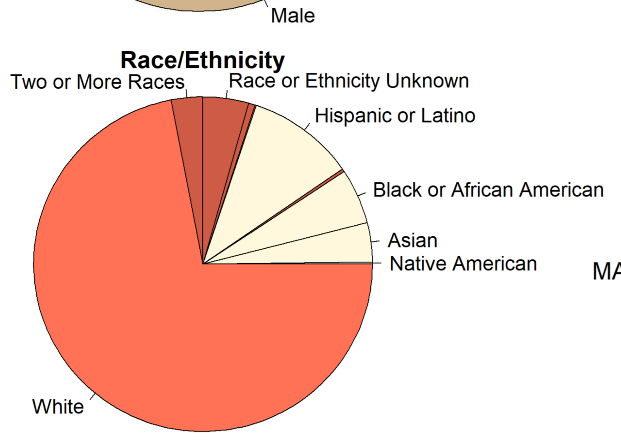


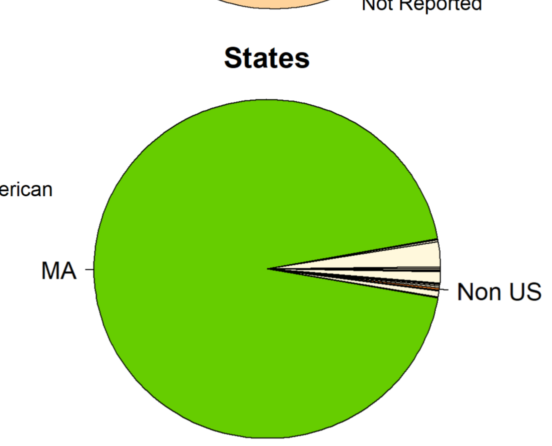


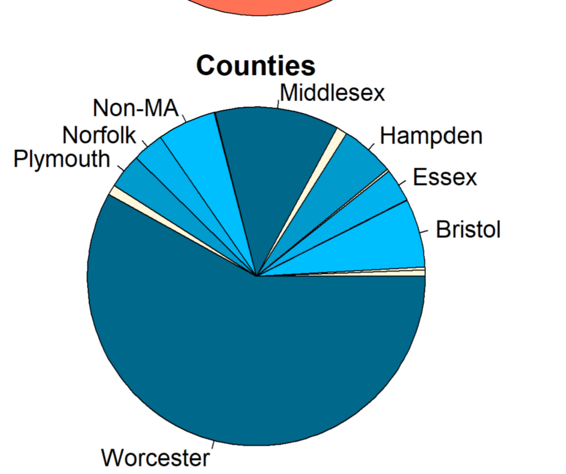
Participant Demographic Information

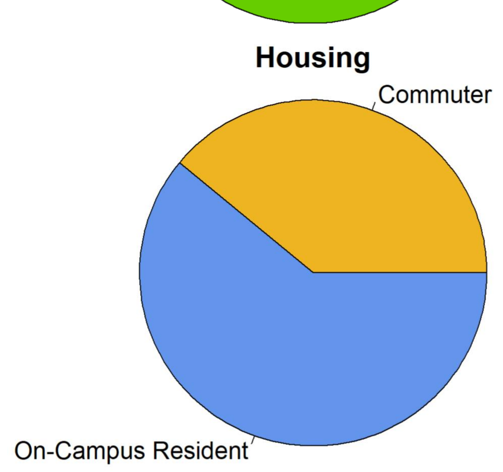












Research Questions

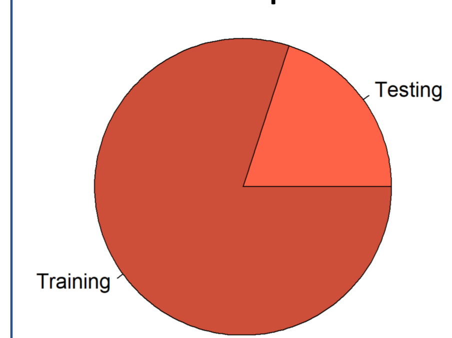
1. Can we build accurate models to predict students' cumulative GPA both numerically and categorically?
2. What factors most accurately predict students' final cumulative GPA?

Methods

For all analysis, data was split into training and testing groups and for all classification, cumulative GPA was divided into three intervals: low, mid, and high.

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

80% Split



Training – randomly selected 80% of the data set used to build the model

Testing – randomly selected 20% of the data set used to test the model for accuracy

Rule-Based Classifiers:

PART

Forms rules off of **all** variables to create a model. Rules are of the 'if, else…' format, **many** rules can result from this algorithm.

OneR

Generates one rule for each predictive variable and then chooses the rule with the least error as its overarching rule.

ZeroR

Does not take any variables into account; ZeroR determines a 'rule' based on the distribution of data. The class with the most instances is assigned to any new instances being classified.

Results

```
=== Confusion Matrix ===

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

```
=== Confusion Matrix ===

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

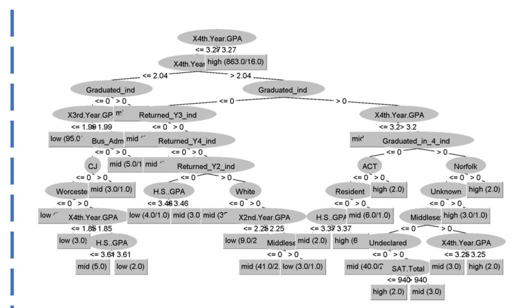
```
=== Confusion Matrix ===

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

Decision Tree Algorithms:

J48

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.

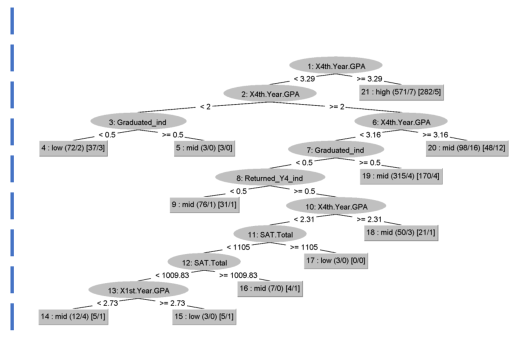


```
=== Confusion Matrix ===

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

REPTree

Generates a decision tree, 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.



```
=== Confusion Matrix ===

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

Random Forest

Consists 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.

```
=== Confusion Matrix ===

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

K-Nearest Neighbors

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.

```
=== Confusion Matrix ===

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

Naïve Bayes

Uses Bayesian statistics to find the likelihood that an instance belongs in each class using conditional probability and Bayes' Rule. The probability for each class is calculated, then the class with the highest likelihood is assigned to that instance.

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

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

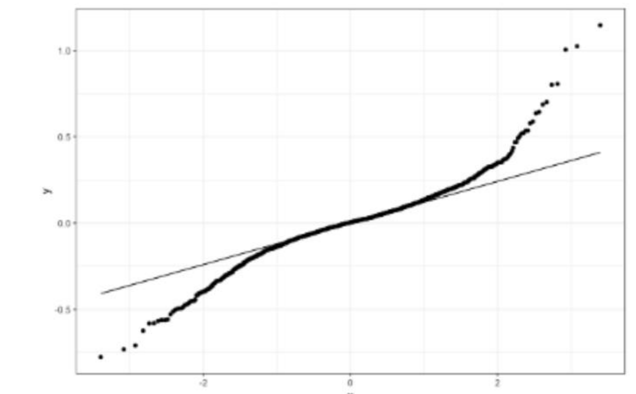
```
=== Confusion Matrix ===

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

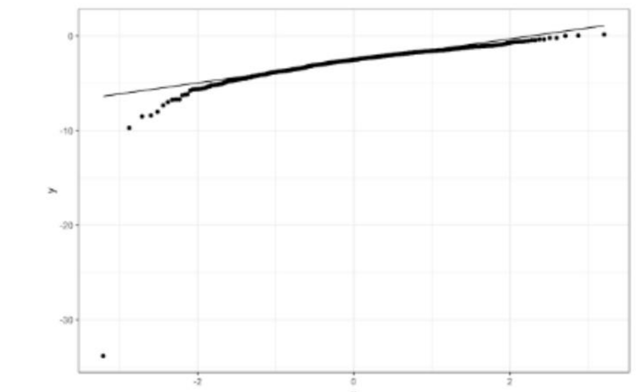
Linear Regression

Built a linear regression model using R, which considers every variable given and develops a model using the subset of those variables that describes the data with the highest accuracy. Weights are applied to each variable to predict cumulative GPA.

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



(a) Original QQ-plot of residuals



(b) QQ-plot of log of residuals

Accuracy Rates

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

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

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

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

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

Conclusions

We were able to build accurate models used to predict cumulative GPAs of Worcester State students.

Our linear regression model was the best model with an accuracy rate of 97.01%.

The variables in the linear regression model contribute the most to cumulative GPA. White, race and/or ethnicity Unknown, 1st year GPA, 4th year GPA, Returned Y3, Returned Y4, Graduated in 4, and Graduated in 8 are some of the most statistically significant variables in the model.

The results of this study can be reported to faculty, administration, and students to give them the resources to improve student success and provide support to struggling students that can be identified as at-risk of failure due to the underlying factors pinpointed in this study.