Using Machine Learning to Maximize First-Generation Student Success A Contribution to the Mission of Aiding the Underserved

Mustafa Emre Yesilyurt

San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_projects

Part of the Artificial Intelligence and Robotics Commons

Recommended Citation
DOI: https://doi.org/10.31979/etd.bspf-nqv9
https://scholarworks.sjsu.edu/etd_projects/1198

This Master's Project is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Projects by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.
Using Machine Learning to Maximize First-Generation Student Success

A Contribution to the Mission of Aiding the Underserved

A Project

Presented to

The Faculty of the Department of Computer Science

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Mustafa Emre Yesilyurt

September 2022
The Designated Project Committee Approves the Project Titled

Using Machine Learning to Maximize First-Generation Student Success

A Contribution to the Mission of Aiding the Underserved

by

Mustafa Emre Yesilyurt

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSÉ STATE UNIVERSITY

September 2022

Dr. Teng Moh Department of Computer Science
Dr. Melody Moh Department of Computer Science
Dr. Elaine Collins Department of Chemistry
ABSTRACT

Using Machine Learning to Maximize First-Generation Student Success
A Contribution to the Mission of Aiding the Underserved

by Mustafa Emre Yesilyurt

The Leadership and Career Accelerator (UNVS 101) is a course offered at San José State University (SJSU) designed to hone industry skills in and provide support to students of underserved backgrounds. The main goal of this study is to determine which features are most significant to identifying the students at risk of failing the course. This will allow faculty to better focus data collection efforts and facilitate an increase in classifier accuracy. The data came as three distinct sets (sources). One contained features describing student demographics and academic history, another described the students’ experience in the course, and a third showed the students’ readiness for the industry. The latter two were collected as survey responses from students. To score these features, three different machine learning (ML) techniques were used in two different schemes, although one of the schemes (principal component analysis a.k.a. PCA) was found to be ill-suited for this data set and for feature selection as a whole. The remaining scheme showed the survey features to be the most significant. ML models corroborated findings by testing specific feature subsets. Some overfitting was observed due to the data’s small size, but accuracies were consistently higher in models that were trained with the survey features. Additionally, a method employing clustering was developed that provided even more evidence of the survey features’ significance. This study ultimately proves that the UNVS 101 students’ survey responses are decidedly more important than their demographics and background when it comes to identifying their at-risk status. With continued diligence in data collection efforts, this study can be revisited using an expanded data
set that will prove instrumental in more accurately identifying at-risk students.
# TABLE OF CONTENTS

## CHAPTER

1. **Introduction** .................................................. 1  
   1.1 Background .................................................. 1  
   1.2 Related Study .............................................. 2  
   1.3 Contributions and Article Guide ..................... 4  

2. **Methodology** .................................................. 9  
   2.1 Features and Ranking Scheme Design .............. 9  
   2.2 Ranking Corroboration Using ML Classifiers ...... 12  
      2.2.1 Truth Label Interpretation Scheme .......... 12  
      2.2.2 Model Choice and Accuracy Measurement ... 13  
   2.3 Ranking Corroboration Using Clustering .......... 14  
      2.3.1 Visualization with Dendrograms .......... 15  
      2.3.2 Individual Records Case Study ............ 16  
      2.3.3 Confusion Matrices ....................... 17  

3. **Results** ..................................................... 19  
   3.1 Feature Ranking ...................................... 19  
   3.2 Model Accuracy ..................................... 24  
   3.3 Clustering Efficacy .................................. 27  
      3.3.1 Visualization with Dendrograms ......... 27  
      3.3.2 Individual Records Case Study .......... 34  
      3.3.3 Confusion Matrices ....................... 55
4 Discussion ................................................................. 84
  4.1 Making a Statement with ML Classifiers ................. 84
  4.2 The Inadequacy of PCA ........................................ 85
  4.3 Confirmation with Clustering ............................... 87
    4.3.1 Visualization: Exploring Visible Feature Significance ... 87
    4.3.2 Measuring Clustering Quality .......................... 89
    4.3.3 Case Study: Elucidating Feature Shortcomings ..... 89
    4.3.4 Confusion Matrices: An Urgent Need for More Data 90
  4.4 A Reflection ..................................................... 93

5 Conclusion and Future Work ....................................... 96

LIST OF REFERENCES .................................................. 99

APPENDIX
CHAPTER 1

Introduction

1.1 Background

College students who are first-generation (students whose parents did not attend college) and who come from low-income backgrounds have been found to struggle with completing their degrees. According to the U.S. Department of Education, only twenty percent of first-generation students nationwide complete their Bachelor’s program within ten years of their sophomore year of high school; this stands in contrast to continuing-generation students’ forty-two percent statistic [4]. Of those first-generation students that graduate, however, many still have trouble securing employment positions for which their degrees would render them qualified. This problem has been attributed primarily to present support systems being ineffective or even nonexistent as well as the lacking of skills and networks that are necessary to be successful in the industry [1, 2, 3, 4]. UNVS 101 was developed to address both of these problems. Note that while not all students who take the course are specifically first-generation or hail from underserved communities, UNVS 101 was designed specifically with these individuals (and the needs that, statistically, they generally share) in mind. To address these needs, students are provided with an environment that helps build work ethic, assignments that hone career skills such as networking, interviewing, and resume building, and administrative support to address any inter- or intrapersonal issues that may arise. Of course, there are students who do not succeed in the course despite the continued efforts and vigilance from faculty. In order to reach and guide these students to a more successful path, they need to be identified. Thus, the overall goal of this study is to utilize both data about the students and UNVS 101 data to highlight the features most critical to indicating whether or not a student will succeed in the course. To confirm this method’s findings, machine
learning models (or “predictors”) that could identify which students were in danger of not succeeding in the course would be developed; it is expected that, depending on which features are utilized, the models’ accuracies would differ. This would provide sufficient evidence to the argument of which features were most significant to at-risk student identification. Secondary to the gathering of evidence, another reason for the models’ development is for use by UNVS 101 staff as a tool to identify at-risk students during the course of the semester.

1.2 Related Study

Many studies have been done to try to identify the struggles of first-generation and/or underserved students. Two papers that were of particular interest include Redford and Hoyer’s “First-Generation and Continuing-Generation College Students: A Comparison of High School and Postsecondary Experiences” and Conefrey’s “Supporting First-Generation Students’ Adjustment to College with High-Impact Practices.” Between just these entries, a clear picture of the need for this research and why improving the students’ experience with this course is established.

Redford and Hoyer’s research attempted to further understand the long-term differences between first-generation and continuing-generation college students by recording the ethnic and financial backgrounds, personal expectations, and academic performances of the students at a specific point (the group’s sophomore year of high school, 2002) and noting their employment standing or current schooling (or reasons for leaving it) ten years later (2012) [3]. This data may not be very timely, but there is an emphasis on the idea that performance within the school environment is greatly affected by problems the students have outside the school, the ideas therein being that support staff are limited in how greatly they can address these problems at school and that a deep understanding of the specific problems students face is how they can
be reached.
Ultimately, while the article does not provide many details about those problems, the statistics surrounding them are highly expounded upon; it shows irrefutable results dictating that continuing-generation students are uniformly more advantaged, hold higher levels of educational attainment, and are not as greatly afflicted by the same problems leading to dropping out as first-generation students.
Conefrey's paper, conversely, focuses less on collecting and comparing statistics, instead favoring the employment of psychiatric study to understand the problems plaguing first-generation students. The research here shines light upon why first-generation students hold lower degree completion and retention rates, finding that a primary reason was a lack of guidance and support from authority figures [4]. Additionally, a feeling of unfamiliarity due to cultural and class-based norms was cited. This study also openly acknowledges the emotional problems arising from financial concerns (effectively lending credibility to the previous paper's statistical analysis), lower self-esteem, and underdeveloped skills with regard to work ethic and time management. The paper goes on to detail the benefits of certain practices, many of which are found in UNVS 101: seminars, writing-intensive courses, collaborative assignments, and more. All of this is in service to the notion that students' expressiveness and connectivity with peers and support staff is the gateway to better understanding their specific problems and addressing them accordingly.
There have also been other studies that employ machine learning techniques to try and identify troubled students in a college environment. For instance, Prenkaj et al. took a very mathematical approach to finding models that could predict which students would be dropping out [5]. The article itself is both very sprawling and in-depth, covering a wide breadth of topics such as a lengthy study of existing literature and techniques, extending the current models to different contexts such as online learning
environments, the presence (or lack thereof) of different data sources and features, comparisons of the strengths and weaknesses of the different techniques, and how those techniques are actually employed with the data. For as much detail as is present in the paper, however, it does not explore any of the features’ significance relative to each other, much less provide a specific ranking.

1.3 Contributions and Article Guide

The sections and subsections of the article below are as follows:

1. Methodology: This section covers the logic and thought processes that went into how the experiments were conceived and conducted.

   (a) Features and Ranking Scheme Design: This is where the features are listed by order of source and the different ranking methods are introduced. Basic technical details and parameters are also provided.

   (b) Ranking Corroboration Using ML Classifiers: With the feature ranking schemes having generated results, further evidence is needed to make the findings more conclusive. Supervised learning classifiers are created with standard libraries to compare accuracies based on different feature vectors.

     i. Truth Label Interpretation Schemes: The first step was deciding to use the final course grade as the target label; the next was to use it effectively and efficiently. This subsection describes the reasoning for why and how the grade data was interpreted the way it was.

     ii. Model Choice and Accuracy Measurement: Here is where the fundamental technologies, selected techniques, and model testing schemes are described.
(c) **Ranking Corroboration Using Clustering:** In contrast to the aforementioned ML models, clustering (an unsupervised learning technique) is employed in three different schemes in order to attempt to reach the same results through different means. Also explored is what parameters are considered (or controlled) in these experiments.

i. **Visualization with Dendrograms:** Dendrograms are a primary method of visualizing clustering patterns. In this study, they are generated using the Python "scipy" library’s "hierarchy" module (contained within the module "cluster") and are only dependent on the features’ weight configuration. Also explored is the distance function that enables the weighting.

ii. **Individual Records Case Study:** The design of this scheme adds the grade threshold parameter from the supervised learning schemes. It allows for the questioning of whether or not some feature vectors should be considered of good quality by measuring the placement of like records in the dendrograms.

iii. **Confusion Matrices:** This last scheme was partially designed to be a tool for administrative staff to use to more rapidly identify at-risk students. The clustering trial (dependent on the weight configuration) is stopped prematurely using a custom-written clustering algorithm once a minimum cluster size parameter is reached. The records in the largest cluster are labeled "positive" and the rest are considered "negative". The true/false positive/negative counts and rates are dependent on the grade threshold.

2. **Results:** This section contains several tables and figures detailing the results of
the experiments described in "Methodology" (Chapter 2).

(a) **Feature Ranking:** Rankings based on the classifiers' feature importance scores are seen in three separate bar graphs. The graphs differ in what subset of features they focus on. PCA results are collected in a single table.

(b) **Model Accuracy:** This subsection goes into the different models' accuracy test results as detailed in several tables. The tables are designed to mirror the bar graphs' style of difference in feature vector.

(c) **Clustering Efficacy:** The threshold for what values of "grade" cause a student to be designated as "at-risk", the size of the subset of students assumed to be passing, and the feature set and associated weights all act as parameters with this technique.

1. **Visualization with Dendrograms:** The figures here show the completed clustering patterns dependent only on the features' weight configuration.

2. **Individual Records Case Study:** This scheme more closely examines individual records seen in the same dendrograms from Section 3.3.1. It merely highlights the indices on the figures that are unrecognizable (as presented in this paper) given the small font size.

3. **Confusion Matrices:** Figures of confusion matrices detailing the true/false positive/negative counts and rates are presented here. They are generated using the Python "sklearn" library's "metrics" module.

3. **Discussion:** With the experiments completed, the results are subjected to analysis and interpretation in this section. Conclusions are drawn about the implications of which features are the most significant. Why the models and
clustering algorithm perform as they do is also touched upon.

(a) **Making a Statement with ML Classifiers:** This section references the tables and bar graphs of Sections 3.1 and 3.2 to provide arguments for why some features are significant, why others are abjectly harmful, and what improvements can be made.

(b) **The Inadequacy of PCA:** The ways in which PCA were ultimately unsuitable for this project are explored here.

(c) **Confirmation with Clustering:** The three schemes employed are further analyzed in order to examine how they provide corroborating evidence for Section 3.1’s results.

   i. **Visualization: Exploring Visible Feature Significance:** This subsection acknowledges the limitations of clustering visualization but details the ways in which it is helpful to this study.

   ii. **Measuring Clustering Quality:** While noting where visualization comes up short, this subsection highlights the need for empirical evidence and how the following two schemes provide it.

   iii. **Case Study: Elucidating Feature Shortcomings:** Here, an attempt at parsing out feature-based patterns in the grouping of "positive" and "negative" cases is attempted. Also explored is why such patterns do or do not exist and how those conclusions compare with those of the ML classifiers.

   iv. **Confusion Matrices: An Urgent Need for More Data:** In this penultimate subsection, the confusion matrix experiments (both the table and figures) are cited in explaining the need for a larger data set.
and how this connects to earlier supervised-learning schemes. Notes for future iterations of this project are also mentioned.

(d) **A Reflection:** Finally, the Discussion closes with a rumination on the humanity at the heart of this project and what pursuing it meant to its contributors.

The fundamental purpose of this study is to highlight the most important features in the provided data set. These results are then backed up with hard evidence generated by creating machine learning models that essentially test the efficacy of each primary subset of features: All Sources vs Source 1 (demographics, academic and personal history) vs Sources 2 and 3 (survey responses regarding performance in UNVS 101 and industry readiness).
CHAPTER 2
Methodology

Technology used in this study include Python, Jupyter Notebook (a Python development environment), Microsoft Excel, and MySQL. Within Python, the primary libraries used include “numpy” and “pandas” for data reading, cleaning, displaying, and manipulation as well as “sklearn” and “statistics” for ML technique implementations and statistical computations, respectively. Microsoft Excel and MySQL were used much earlier in the study; Excel was employed to initially view the data and conduct higher-level data manipulation and wheedling-out of unusable records. MySQL was used in the cleaning of Source 2 data (weekly survey responses) by collapsing multiple rows into single records with average scores grouped by the anonymized student identifier. Due to many records not having the survey features completed, the number of usable records available totaled to only 79.

2.1 Features and Ranking Scheme Design

Table 1 details the features’ names, parent source, overall index, and data type. In Source 1, it can be seen that the demographic features (numbered 1 through 10) are all of the binary type, meaning they are either of value 0 or 1 with 1 indicating that the student identifies as being of that listed group. Features 11 and 12 are interpreted from qualitative (written-word) data by creating a response-to-number mapping of possible values as the ML techniques required quantitative (numerical) data. Feature 11 details the individual’s gender identity which was limited to only a few possible values, but feature 12, the student’s major, showcased much more variety. The students’ grade point average (GPA) is the most complex of the bunch as a decimal value. Their final course grades (feature number 27 for ML model input simplicity) also belongs in Source 1 and acts as the truth label for the classification
schemes described in the next section.

The Source 2 data (features 15 through 18) covers the weekly survey responses from students and the number of weekly collaborative lab sessions they attended throughout the semester. There are only three questions on the survey: how likely the student is to use what they learned that day, how close they feel to their peers, and how engaged they were that session. These features are listed as being decimal values because they are, in fact, averages gathered by collapsing multiple rows; the actual student responses are on a scale of 1 to 10, but since different students attend a different number of total weekly sessions, aggregating them in this fashion and adding a counter for the number of sessions attended (“num_labs_attended”) was the compromise.

Table 1: Features in the data set used in this study.

<table>
<thead>
<tr>
<th>Source Number</th>
<th>Feature Number</th>
<th>Feature Name</th>
<th>Notes on Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>person_of_color</td>
<td>Binary value</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>white</td>
<td>Binary value</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>african_american</td>
<td>Binary value</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>asian_american</td>
<td>Binary value</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>latino</td>
<td>Binary value</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>native_alaskan</td>
<td>Binary value</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>native_american</td>
<td>Binary value</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>native_hawaiian</td>
<td>Binary value</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>pacific_islander</td>
<td>Binary value</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>multi_ethnic</td>
<td>Binary value</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>gender_identity</td>
<td>Qualitative; 5 known values</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>major_braven</td>
<td>Qualitative; 57 known values</td>
</tr>
<tr>
<td>1</td>
<td>13</td>
<td>num_intern</td>
<td>Integer value between 0 and 7</td>
</tr>
<tr>
<td>1</td>
<td>14</td>
<td>gpa</td>
<td>Decimal value</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>avg_likely_use</td>
<td>Decimal value</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>avg_close_feel</td>
<td>Decimal value</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>avg_engagement</td>
<td>Decimal value</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>num_labs_attended</td>
<td>Integer value between 1 and 14</td>
</tr>
<tr>
<td>3</td>
<td>19</td>
<td>ability_certainty</td>
<td>Integer value between 1 and 7</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>confidence_what_it_takes</td>
<td>Integer value between 1 and 7</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>search_certainty</td>
<td>Integer value between 1 and 7</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>career_decision_confidence</td>
<td>Integer value between 1 and 7</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>can_make_career</td>
<td>Integer value between 1 and 7</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>grow_professionally_confidence</td>
<td>Integer value between 1 and 7</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>capable_dealing_problems</td>
<td>Integer value between 1 and 7</td>
</tr>
<tr>
<td>3</td>
<td>26</td>
<td>performance_confidence</td>
<td>Integer value between 1 and 7</td>
</tr>
<tr>
<td>1</td>
<td>27</td>
<td>grade</td>
<td>Decimal value (target label)</td>
</tr>
</tbody>
</table>
The features of Source 3 (numbered 19 to 26) cover student responses did not require any aggregation because there was only one record per student. Thus, they were able to remain on the integer scale that the students themselves filled out. The features themselves are as follows: how certain the student feels in their ability to get the job they want, how confident they are that they have what it takes to get a good job, how certain they are that their job search will be successful, how confident they are that their career decisions will be good, how strongly they believe that they can do what they need to make their career, how confident they are in their ability to grow and improve professionally, how capable they are in dealing with problems that come up in their career, and how confident they are in their ability to perform well at work. There are more questions on the survey that were added later, such as the students’ ability to use tools such as LinkedIn and other questions related to job search readiness, but there were not enough recorded responses to allow the inclusion of those features.

The scoring was done in Python by building a classifier (either the DecisionTreeClassifier object or RandomForestClassifier object, both from “sklearn”) without any added parameters, fitting the data to it, and saving the results of the “feature_importances_” attribute to a list. The scores are then matched to their respective features and displayed on bar graphs. Both techniques’ scoring outputs share the same graphs so that the bars for each technique are side-by-side, making the comparison of their assigned scores easier. The features themselves were reordered for the graphs such that the random forest-based scores were listed in descending order. This was done because, as is also displayed at the bottom of the graph, the random forests scores have a lower standard deviation than the decision tree’s scores, meaning the former has the more reliable results. Across every bar graph with the data set used, the random forest always had the lower standard deviation.
Feature scoring was not limited just to these two ML techniques, however. Principal component analysis (PCA) was considered much earlier in the project’s development as a viable tool in understanding which features were more significant. When it came to employing the technique, a table was generated that details the importance of each individual feature by noting the total information loss when a specific feature was left out; the larger the loss, the more important that feature is. This loss analysis was observed for all possible numbers of principal components (i.e. two to twenty-five), but only the results from two to four are shown in the table.

For both of these ranking schemes, three figures/tables were generated: the first shows the analysis when all features are present, the second with only the non-survey Source 1 columns, and the third with only the Source 2 and 3 survey response features.

2.2 Ranking Corroboration Using ML Classifiers

The ranking of the features by score or information loss is the main point of this study, but confirmation of these results would be ideal. Thus, the idea to train classifiers with subsets of the large feature vector was conceived; by controlling which features were included in the training data set, the resultant model accuracy would indicate the significance of the features included relative to the ones left out. To do this kind of classification, however, requires the usage of a truth label.

2.2.1 Truth Label Interpretation Scheme

As the overarching purpose of the data is to identify whether or not a given student is at risk of failing UNVS 101, it is sensible for the classifiers used to corroborate the feature rankings to generate a binary output. This would mean, then, that the final course grade feature (“grade”) in the data set would need to be reduced from a decimal value into a binary one. Depending on the grade alone, a student’s record would be
judged as successful or failing in order to train on the resultant record with a binary label.

Rather than set a permanent threshold of what grades should be considered passing or failing, it was decided to make this threshold a variable to experiment with and observe resultant fluctuations in accuracy. The way the threshold scheme works is decidedly simple: any grade that is less than the threshold is reduced to 0 and all other grades are set to 1. Many tests were run with different thresholds, leading to intriguing findings regarding test accuracy with respect to threshold changes. It was also suggested that a non-binary multi-class partition scheme be devised as that may give further insight. This partition scheme separates the students into five distinct groups based on their value for “grade”: “definitely fail”, “probably fail”, “neutral”, “probably pass”, “definitely pass”. Depending on the group to which the given record belongs, “grade” is replaced with a number (zero to four, respectively). While giving different results from the binary threshold scheme, it is believed that both techniques suffered from the same problems with the data set.

2.2.2 Model Choice and Accuracy Measurement

The ML techniques used for the classifiers were decision trees (DTs), random forests (RFs), support vector machines (SVMs), and artificial neural networks (ANNs). All of these techniques were selected specifically for the unique perspectives each technique provides. SVMs were employed precisely because of their high accuracy rates and resistance to overfitting while neural networks, or rather the multi-layer perceptron (MLP), were a sound choice to allow for greater parameter manipulation and freedom in experimenting with performance maximization. In addition to mirroring the techniques used in the feature ranking bar graph scheme, decision trees and random forests were both used in order to understand the effect of ensemble
learning on this data set.

To be able to rapidly test these models simultaneously while also making adjustments to parameters, a helper function for each technique was designed that took in the training and test data, built the classifier from its “sklearn” implementation with the referenced parameters, made the predictions on the test data, and returned the predictions' accuracy as computed by the “metrics” module found in “sklearn”. These function call-generated accuracies are then stored in lists which, in turn, are populated by iterations through the different grade thresholds or partition sets being tested. 5-fold cross-validation was also employed during testing with each model’s accuracy being averaged across all the folds. What is observed in the Results section is a final 10-fold cross validation scheme’s output table. A total of six tables are present: the first two tables are the binary and partition threshold scheme results for all features, respectively, the next two for just the Source 1 features, and the last two for the Source 2 and 3 features.

2.3 Ranking Corroboration Using Clustering

In employing an unsupervised learning technique as versatile as clustering, it is important to attempt to utilize as many facets of the technique as possible in order to find corroborating evidence. With this particular section, three different schemes are designed, each utilizing a different strength of the clustering technique: visualization of complete clustering trials, tracking of individual records’ clustering patterns, and confusion matrices derived from clustering trials that are prematurely terminated. Different parameters are also explored across these three schemes. The most critical of these variables is the weight configuration of the feature vector, appearing in every one of the schemes. The grade threshold that was introduced in the previous section
for the supervised learning ML classifiers is relevant again but only in the confusion matrix scheme. The last parameter to consider is a minimum cluster size (or "cluster size threshold") that dictates the termination point in a clustering trial, which is also only utilized with the confusion matrices. These variables are further explained alongside their respective schemes.

2.3.1 Visualization with Dendrograms

A dendrogram is an often-relied upon medium of visualizing clustering patterns and sequences. Dendrograms are graphs that allow the viewer to see both the order in which records are clustered together (with the records’ index being visible along the x-axis) as well as the distance between the two clusters/records that are being merged (seen along the y-axis). Several dendrograms are displayed in this study in order to demonstrate how the clustering algorithms ultimately behave. It should be noted, though, that these dendrograms alone are not enough to gauge the clustering trials’ effectiveness; as implemented in Python’s "scipy" library, dendrograms are generated as a result of merging records until only one cluster is present with no method of interrupting them, meaning they are not useful in gathering metrics. However, by comparing the graphs, it is possible to witness the overall effects of the feature selection and weighting methods detailed later in this section. To actually quantify these effects requires a different approach (calculating the confusion matrices) that necessitates being able to step through the clustering process and stop prematurely. The precise nature of that algorithm is detailed later on with the confusion matrix scheme in Section 2.3.3.

What is more relevant to the visualization aspect is the distance function created in order to apply weighting. In this study, the weighted Euclidean distance function was the function of choice; rather than subject each of the clustering trials to employ the
same uniform weight distribution to each of the features in the data set, this method allows the weight assigned to each feature to be parameterized, which, in turn, allows the tuning of the weights such that they align with the features’ significance ranking as observed later in Section 3.1. If the results are better with such weight distribution, then those results can act as further corroborating evidence of the earlier findings. Additionally, out of all the parameters detailed in this section, the only parameter involved in generating the dendrograms are the weights. Other distance functions are available (Chebyshev, Mahalanobis, Manhattan, Minkowski, etc.) and could be further experimented with in the future, but, just as with the ML classifiers of the previous section, clustering performance enhancement is not the primary object of this study. Weighted Euclidean distance performs well for the study’s purposes, is not particularly challenging to implement, and would do exactly what is required in order to find evidence proving the feature significance ranking to be accurate.

2.3.2 Individual Records Case Study

The case study explored in this project requires exactly one more parameters beyond the feature weights, and that is the grade threshold. In this scheme, records that are known to be "positive" or "negative" based on a specific grade threshold value. Through examination of the data set, it was determined that in order to better showcase the distance function’s ability to distinguish the records, a more sizable number of records that could be accepted as being "negative" would be required, so lower thresholds could not reliably be used. Different grade thresholds (0.78, 0.8, 0.875) were also tested, but 0.8 was the only value that generated enough negatives to properly test the cluster size threshold and feature weighting while still maintaining the integrity and distinctiveness of the records (especially ones with grades close to the threshold). Therefore, while the results based on the 0.8 threshold are emphasized
the most, the other thresholds’ results will also be featured in tabular format so as to show their ineffectiveness when employed with this particular data set. In practice, this scheme is really just an extension of the visualizations generated in the previous one. The difference is that the changes in clustering patterns of the actual records across the different weight configurations can be observed here. This in turn will allow the examination of whether or not records that are clustered as a result of the change in weighting should even be clustered at all. The "positive" records that are tracked were specifically selected due to their similarity looking only at their values in the data set. The "negatives" were chosen because they were the only "negatives" available in the set given the 0.8 grade threshold.

2.3.3 Confusion Matrices

As established previously, this final scheme of generating confusion matrices necessitates the premature termination of clustering trials. Thus, it introduces the final parameter to consider: the minimum cluster size (or "cluster size threshold") that dictates the termination point in the clustering trial. In a given trial of this scheme, once there is a cluster that is at least as large as the cluster size threshold, the trial terminates; the records contained within that largest cluster are then labeled "positive" (meaning these students are not at risk of failing) with every other record, regardless of how or if they are clustered, being considered "negative" (that is, at risk of failing). This positive/negative labelling allows calculation of the confusion matrices; the true negative, false negative, true positive, and false positive counts and rates are shown in both figures as well as a table that details the same information in a more structured and readable manner. Note, however, that the figures are specific visualizations of the confusion matrices and do not visualize the clustering trials themselves.

To create a clustering algorithm that could stop premature, the decision of what type
to use needed to be made. The choice of clustering algorithm type was single-linkage, meaning, when updating distances after a merge, the new distance between two clusters is the shortest distance between their respectively contained records; it is because the clustering algorithm is single-linkage that no recalculation of distances between clusters or records is necessary, significantly reducing the time complexity of running a clustering trial.

Also recall that a secondary priority of this project is to provide a method instructors and administrative staff can employ to identify at-risk students as soon as possible. This can be accomplished, to a degree, utilizing the same process of prematurely stopping the clustering process and labeling outliers to the largest cluster as "negatives". It is expected that, given the size of the set and small number of students who performed poorly in the course, there will be a high true positive rate and a low true negative rate. Finding a weight configuration that maximizes both will be helpful to instructors that would be trying to prune the genuine "positives" from the "negative"-labeled records in search of the at-risk students; the student profiles labeled "positive" can be set aside so the "negatives" can be given a closer look to determine at-risk status. This will, at the very least, reduce the number of manual checks necessary and will hopefully prove a valuable tool if and when the course is being taught to more classes and, thus, a much larger number of students.
CHAPTER 3

Results

3.1 Feature Ranking

Figures 1, 2, and 3 showcase the ranking of features using bar graphs that compare the feature scores from the default RF and DT classifiers. In Figure 1, which compares all the features simultaneously, the RF classifier determines that the students’ GPA is the most important with the Source 2 features (the three averages and, with a lower score, “num_labs_attended”) following very closely behind. The DT results disagree with many of these rankings, however; while GPA is still very significant, it comes second to “avg_close_feel”, both of which stand head and shoulders above the remaining Source 2 features. DT emphasizes some of the Source 3 features such as “career_decision_confidence” and “can_make_career” more than those from Source 2.

![Feature Scores (Trained With All Features - 79)](image)

Figure 1: Feature scores based on models trained with all available features (1-26).
Aside from GPA, the students’ major is also highlighted by DT and has a higher score than in the RF results. What the two types agree on, though, is the insignificance of a majority of the Source 1 features, especially the binary demographics columns. This agreement carries over into Figure 2 where only the Source 1 features are present. The demographics are still the least significant features for both RF and DT while GPA and major rank the highest. Notice how the standard deviations listed in this figure are much higher than those of the previous one.

![Feature Scores](image)

**Figure 2:** Feature scores based on models trained with only non-survey features (1-14).

With Figure 3, there is some disagreement between RF and DT, but what is more interesting is that, while the RF rankings are mostly the same outside of a few of the similarly-scoring columns swapping places, the DT scores here conflict somewhat drastically with those found in Figure 1. Where “can_make_career” was among the
more highly-rated by DT in Figure 1, it is now the absolute lowest; conversely, “avg_engagement” had only a middling ranking before but here is the highest scoring out of the DT scores by a fairly large margin compared to those found in the RF scores. It is not surprising, then, that the RF standard deviation continues to be lower than the DT’s (although the difference between them is actually lower than in Figure 1).

Figure 3: Feature scores based on models trained with only survey features (15-26).

Tables 2, 3, and 4 detail the PCA analysis results for all features, Source 1 features, and Source 2 and 3 features, respectively. There is a considerable amount of disagreement between the findings listed in these tables and those found in Figures 1, 2, and 3; in Table 2, a few of the Source 1 demographic features are seen as being of greater significance than most of the Source 2 columns. Furthermore, the top six spots on the list are all taken by Source 3 columns, with the first Source 2 feature taking
the 7th place rank. Notice that “can_make_career” is the ranked the highest; while it was seen as significant according to the DT results in Figure 1 (but not Figure 3, notably) it still did not compare to GPA and “avg_close_feel”, which RF corroborates. This is not the case in Table 2, where “avg_close_feel” ranks 12th, “avg_likely_use” (RF’s favorite feature in Figure 3) is 13th, and the GPA, quite shockingly, is near the very bottom of the ranking at 24th. The students’ GPA and major ranking so low is quite concerning, although it is true that many of the Source 2 and 3 columns rank higher than most of the demographic features.

Table 2: PCA results for all available features.

<table>
<thead>
<tr>
<th>Missing Feature</th>
<th>Info Loss - n_comp = 2</th>
<th>Info Loss - n_comp = 3</th>
<th>Info Loss - n_comp = 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>can_make_career</td>
<td>0.715510</td>
<td>0.618972</td>
<td>0.538735</td>
</tr>
<tr>
<td>career_decision_confidence</td>
<td>0.709334</td>
<td>0.611891</td>
<td>0.531788</td>
</tr>
<tr>
<td>confidence_what_it_takes</td>
<td>0.708104</td>
<td>0.610562</td>
<td>0.532107</td>
</tr>
<tr>
<td>ability_certainty</td>
<td>0.705042</td>
<td>0.607470</td>
<td>0.528951</td>
</tr>
<tr>
<td>grow_professionally_confidence</td>
<td>0.704318</td>
<td>0.607286</td>
<td>0.529746</td>
</tr>
<tr>
<td>capable_dealing_problems</td>
<td>0.702076</td>
<td>0.604583</td>
<td>0.524487</td>
</tr>
<tr>
<td>avg_engagement</td>
<td>0.700771</td>
<td>0.610985</td>
<td>0.532948</td>
</tr>
<tr>
<td>search_certainty</td>
<td>0.699765</td>
<td>0.602310</td>
<td>0.522732</td>
</tr>
<tr>
<td>native_alaskan</td>
<td>0.697511</td>
<td>0.603997</td>
<td>0.527110</td>
</tr>
<tr>
<td>native_american</td>
<td>0.697511</td>
<td>0.603997</td>
<td>0.527110</td>
</tr>
<tr>
<td>performance_confidence</td>
<td>0.697034</td>
<td>0.601512</td>
<td>0.523698</td>
</tr>
<tr>
<td>avg_close_feel</td>
<td>0.694889</td>
<td>0.607979</td>
<td>0.530171</td>
</tr>
<tr>
<td>avg_likely_use</td>
<td>0.693146</td>
<td>0.611477</td>
<td>0.531481</td>
</tr>
<tr>
<td>native_hawaiian</td>
<td>0.689946</td>
<td>0.592549</td>
<td>0.512369</td>
</tr>
<tr>
<td>num_labs_attended</td>
<td>0.688345</td>
<td>0.590764</td>
<td>0.510743</td>
</tr>
<tr>
<td>african_american</td>
<td>0.687107</td>
<td>0.591050</td>
<td>0.509000</td>
</tr>
<tr>
<td>multi_ethnic</td>
<td>0.686989</td>
<td>0.589789</td>
<td>0.50754</td>
</tr>
<tr>
<td>num_intern</td>
<td>0.685749</td>
<td>0.596123</td>
<td>0.516675</td>
</tr>
<tr>
<td>white</td>
<td>0.685675</td>
<td>0.591877</td>
<td>0.512770</td>
</tr>
<tr>
<td>person_of_color</td>
<td>0.685316</td>
<td>0.589046</td>
<td>0.501029</td>
</tr>
<tr>
<td>pacific_islander</td>
<td>0.685154</td>
<td>0.587747</td>
<td>0.508330</td>
</tr>
<tr>
<td>asian_american</td>
<td>0.685143</td>
<td>0.603426</td>
<td>0.525060</td>
</tr>
<tr>
<td>latino</td>
<td>0.684977</td>
<td>0.593703</td>
<td>0.514735</td>
</tr>
<tr>
<td>gpa</td>
<td>0.684892</td>
<td>0.588795</td>
<td>0.510063</td>
</tr>
<tr>
<td>gender_identity</td>
<td>0.684876</td>
<td>0.591922</td>
<td>0.513128</td>
</tr>
<tr>
<td>major_braven</td>
<td>0.684709</td>
<td>0.587475</td>
<td>0.508801</td>
</tr>
</tbody>
</table>
Table 3: PCA results for only Source 1 features.

<table>
<thead>
<tr>
<th>Missing Feature</th>
<th>Info Loss - n_comp = 2</th>
<th>Info Loss - n_comp = 3</th>
<th>Info Loss - n_comp = 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>asian_american</td>
<td>0.682528</td>
<td>0.551098</td>
<td>0.433315</td>
</tr>
<tr>
<td>native_alaskan</td>
<td>0.676651</td>
<td>0.544569</td>
<td>0.436296</td>
</tr>
<tr>
<td>native_american</td>
<td>0.676651</td>
<td>0.544569</td>
<td>0.436296</td>
</tr>
<tr>
<td>num_intern</td>
<td>0.670478</td>
<td>0.530108</td>
<td>0.413461</td>
</tr>
<tr>
<td>white</td>
<td>0.668935</td>
<td>0.534167</td>
<td>0.421849</td>
</tr>
<tr>
<td>person_of_color</td>
<td>0.659880</td>
<td>0.529186</td>
<td>0.411428</td>
</tr>
<tr>
<td>latino</td>
<td>0.659822</td>
<td>0.521010</td>
<td>0.410341</td>
</tr>
<tr>
<td>native_hawaiian</td>
<td>0.658902</td>
<td>0.533091</td>
<td>0.417153</td>
</tr>
<tr>
<td>multi_ethnic</td>
<td>0.658305</td>
<td>0.515591</td>
<td>0.400530</td>
</tr>
<tr>
<td>gpa</td>
<td>0.655769</td>
<td>0.527983</td>
<td>0.410501</td>
</tr>
<tr>
<td>gender_identity</td>
<td>0.654022</td>
<td>0.529992</td>
<td>0.413255</td>
</tr>
<tr>
<td>african_american</td>
<td>0.652882</td>
<td>0.509184</td>
<td>0.397122</td>
</tr>
<tr>
<td>pacific_islander</td>
<td>0.648254</td>
<td>0.505994</td>
<td>0.388508</td>
</tr>
<tr>
<td>major_braven</td>
<td>0.647266</td>
<td>0.513645</td>
<td>0.395529</td>
</tr>
</tbody>
</table>

Table 4: PCA results for only Source 2 and 3 features.

<table>
<thead>
<tr>
<th>Missing Feature</th>
<th>Info Loss - n_comp = 2</th>
<th>Info Loss - n_comp = 3</th>
<th>Info Loss - n_comp = 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>can_make_career</td>
<td>0.442100</td>
<td>0.305839</td>
<td>0.216025</td>
</tr>
<tr>
<td>avg_engagement</td>
<td>0.433528</td>
<td>0.302593</td>
<td>0.212772</td>
</tr>
<tr>
<td>avg_likely_use</td>
<td>0.428463</td>
<td>0.305737</td>
<td>0.218584</td>
</tr>
<tr>
<td>confidence_what_it_takes</td>
<td>0.428334</td>
<td>0.298839</td>
<td>0.209540</td>
</tr>
<tr>
<td>career_decision_confidence</td>
<td>0.426703</td>
<td>0.290276</td>
<td>0.203864</td>
</tr>
<tr>
<td>avg_close_feel</td>
<td>0.424014</td>
<td>0.292239</td>
<td>0.209453</td>
</tr>
<tr>
<td>ability_certainty</td>
<td>0.420001</td>
<td>0.299904</td>
<td>0.217057</td>
</tr>
<tr>
<td>grow_professionally_confidence</td>
<td>0.419401</td>
<td>0.295169</td>
<td>0.208417</td>
</tr>
<tr>
<td>capable_dealing_problems</td>
<td>0.412977</td>
<td>0.288243</td>
<td>0.203508</td>
</tr>
<tr>
<td>search_certainty</td>
<td>0.410175</td>
<td>0.296274</td>
<td>0.206575</td>
</tr>
<tr>
<td>performance_confidence</td>
<td>0.402292</td>
<td>0.292276</td>
<td>0.202786</td>
</tr>
<tr>
<td>num_labs_attended</td>
<td>0.385690</td>
<td>0.252884</td>
<td>0.200636</td>
</tr>
</tbody>
</table>

This almost total reassignment of significance ranking extends to the remaining tables as well, but more so in Table 3 than Table 4. As in Table 2, the Source 1-only ranking puts the GPA and major near or at the very bottom of the list, which completely flies in the face of Figures 1 and 2. The ordering of the demographic features are mostly the same with only a few rearrangements. Also note that the total information loss in each Table 3 cell is slightly lower than its corresponding cell in Table 2.
3.2 Model Accuracy

Comparing Tables 7 and 8 (which show the models’ results using only Source 1 features) with Tables 5, 6 (all features), 9, and 10 (Sources 2 and 3) demonstrates a clear correlation between Source 2 and 3 feature usage and higher model accuracy with lower standard deviations. Furthermore, the accuracies and standard deviations seen in Table 10 actually indicate better overall models than those found in Table 6. Essentially, this is definitively stating that the models perform more effectively without demographics involved. Table 11 shows the results when all features from Source 1 are removed excluding GPA, major, and the number of internships; they are mostly similar to Table 10’s results with only some fluctuation in accuracy and standard deviation, but it generally gives a higher accuracy.

Of all the tables, only Tables 5 and 9 show any degree of overfitting. The strangest instance of overfitting, which one would expect to take hold starting with a specific threshold and applying for all greater thresholds, is the RF model in Table 5, which overfits for thresholds 0.7 and 0.825 but not the threshold in between or any greater than them. In all the tables’ non-overfitting models, there is a generally-observed degradation in model accuracy. This was observed mostly in the binary threshold scheme as the threshold approached the average grade value of 0.874, although there are some instances of this occurring in the partition schemes’ results. This degradation can mostly be attributed to a kind of clustering of students within a small range of grade values; if a threshold bisects a group of very similar students and labels half of them as “failing”, the classifiers will get confused. Once again, there is a kind of exception to this degradation norm. The MLP classifier tends to fluctuate in its performance rather than consistently degrade. It seems that the presence of Source 2 and 3 features is what causes this variance.

The difference in truth label interpretation scheme is also intriguing, with the partitions
seemingly mitigating the effect of the performance degradation. This is in contrast to the binary scheme which, when not overfitting, shows a mostly steady and decidedly greater decline in accuracy. It also seems that, although the standard deviations are quite high in some instances, the RF models tended to be more reliable than the other techniques’ models. This is to be expected considering the lower standard deviation observed for RF in Figures 1, 2, and 3. That said, the fact they sometimes overfit is concerning; of the model types, the MLP never indicated any possibility of overfitting, although the models’ reliability can be called into question as they sometimes had high standard deviations and were not consistently as low as the RF models’. The overfitting itself can be attributed to the training set being very small; the only way to remedy this problem is to collect more usable data in the future.

Table 5: Results of 10-fold cross-validation for binary threshold scheme using all features.

<table>
<thead>
<tr>
<th>Grade Threshold</th>
<th>Pass/Fail Count</th>
<th>DT Acc.</th>
<th>DT Std Dev.</th>
<th>RF Acc.</th>
<th>RF Std Dev.</th>
<th>SVM Acc.</th>
<th>SVM Std Dev.</th>
<th>MLP Acc.</th>
<th>MLP Std Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.700</td>
<td>77/2</td>
<td>0.973214</td>
<td>0.056626</td>
<td>1.000000</td>
<td>0.000000</td>
<td>1.0</td>
<td>0.0</td>
<td>0.880357</td>
<td>0.198279</td>
</tr>
<tr>
<td>0.750</td>
<td>77/2</td>
<td>0.973214</td>
<td>0.056626</td>
<td>0.987500</td>
<td>0.039528</td>
<td>0.975000</td>
<td>0.039528</td>
<td>0.987500</td>
<td>0.039528</td>
</tr>
<tr>
<td>0.800</td>
<td>71/8</td>
<td>1.000000</td>
<td>0.000000</td>
<td>0.975000</td>
<td>0.052705</td>
<td>1.0</td>
<td>0.0</td>
<td>0.921429</td>
<td>0.112372</td>
</tr>
<tr>
<td>0.825</td>
<td>69/10</td>
<td>1.000000</td>
<td>0.000000</td>
<td>1.000000</td>
<td>0.000000</td>
<td>1.0</td>
<td>0.0</td>
<td>0.848214</td>
<td>0.241384</td>
</tr>
<tr>
<td>0.840</td>
<td>62/17</td>
<td>1.000000</td>
<td>0.000000</td>
<td>0.937500</td>
<td>0.106230</td>
<td>0.937500</td>
<td>0.106230</td>
<td>0.857143</td>
<td>0.241384</td>
</tr>
<tr>
<td>0.850</td>
<td>61/18</td>
<td>1.000000</td>
<td>0.000000</td>
<td>0.960714</td>
<td>0.063442</td>
<td>0.962500</td>
<td>0.086536</td>
<td>0.832143</td>
<td>0.154349</td>
</tr>
</tbody>
</table>

Table 6: Results of 10-fold cross-validation for partition threshold scheme using all features.

<table>
<thead>
<tr>
<th>Partition Set</th>
<th>Count per Partition</th>
<th>DT Acc.</th>
<th>DT Std Dev.</th>
<th>RF Acc.</th>
<th>RF Std Dev.</th>
<th>SVM Acc.</th>
<th>SVM Std Dev.</th>
<th>MLP Acc.</th>
<th>MLP Std Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7, 0.75, 0.775, 0.8</td>
<td>2, 0, 2, 4, 71</td>
<td>0.965205</td>
<td>0.052705</td>
<td>0.900000</td>
<td>0.192916</td>
<td>0.898214</td>
<td>0.192916</td>
<td>0.867857</td>
<td>0.242781</td>
</tr>
<tr>
<td>0.73, 0.775, 0.8, 0.83</td>
<td>2, 2, 4, 4, 67</td>
<td>0.982144</td>
<td>0.086536</td>
<td>0.975000</td>
<td>0.099264</td>
<td>0.987500</td>
<td>0.099264</td>
<td>0.898214</td>
<td>0.224026</td>
</tr>
<tr>
<td>0.68, 0.71, 0.70, 0.81</td>
<td>2, 0, 1, 5, 71</td>
<td>0.960714</td>
<td>0.063442</td>
<td>0.960714</td>
<td>0.063442</td>
<td>0.960714</td>
<td>0.063442</td>
<td>0.832143</td>
<td>0.154349</td>
</tr>
<tr>
<td>0.82, 0.835, 0.85, 0.865</td>
<td>5, 7, 2, 3, 58</td>
<td>0.975000</td>
<td>0.052705</td>
<td>0.900000</td>
<td>0.192916</td>
<td>0.962500</td>
<td>0.063442</td>
<td>0.832143</td>
<td>0.154349</td>
</tr>
</tbody>
</table>

Table 7: Results of 10-fold cross-validation for binary threshold scheme using Source 1 features.

<table>
<thead>
<tr>
<th>Grade Threshold</th>
<th>Pass/Fail Count</th>
<th>DT Acc.</th>
<th>DT Std Dev.</th>
<th>RF Acc.</th>
<th>RF Std Dev.</th>
<th>SVM Acc.</th>
<th>SVM Std Dev.</th>
<th>MLP Acc.</th>
<th>MLP Std Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.700</td>
<td>77/2</td>
<td>0.973214</td>
<td>0.056626</td>
<td>0.973214</td>
<td>0.056626</td>
<td>0.973214</td>
<td>0.056626</td>
<td>0.987500</td>
<td>0.039528</td>
</tr>
<tr>
<td>0.750</td>
<td>77/2</td>
<td>0.973214</td>
<td>0.056626</td>
<td>0.973214</td>
<td>0.056626</td>
<td>0.973214</td>
<td>0.056626</td>
<td>0.973214</td>
<td>0.056626</td>
</tr>
<tr>
<td>0.800</td>
<td>71/8</td>
<td>0.973214</td>
<td>0.056626</td>
<td>0.987500</td>
<td>0.039528</td>
<td>0.987500</td>
<td>0.039528</td>
<td>0.987500</td>
<td>0.039528</td>
</tr>
<tr>
<td>0.825</td>
<td>69/10</td>
<td>0.710714</td>
<td>0.165343</td>
<td>0.835714</td>
<td>0.131977</td>
<td>0.873214</td>
<td>0.131977</td>
<td>0.807143</td>
<td>0.124232</td>
</tr>
<tr>
<td>0.840</td>
<td>62/17</td>
<td>0.619643</td>
<td>0.221129</td>
<td>0.669643</td>
<td>0.160880</td>
<td>0.708214</td>
<td>0.160880</td>
<td>0.708214</td>
<td>0.160880</td>
</tr>
</tbody>
</table>

25
Table 8: Results of 10-fold cross-validation for partition threshold scheme using Source 1 features.

<table>
<thead>
<tr>
<th>Partition Set</th>
<th>Count per Partition</th>
<th>DT Acc.</th>
<th>DT Std Dev.</th>
<th>RF Acc.</th>
<th>RF Std Dev.</th>
<th>SVM Acc.</th>
<th>SVM Std Dev.</th>
<th>MLP Acc.</th>
<th>MLP Std Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7, 0.75, 0.775, 0.8</td>
<td>2, 4, 4, 4</td>
<td>0.948214</td>
<td>0.089266</td>
<td>0.948214</td>
<td>0.089266</td>
<td>0.948214</td>
<td>0.089266</td>
<td>0.948214</td>
<td>0.089266</td>
</tr>
<tr>
<td>0.73, 0.775, 0.8, 0.83</td>
<td>2, 4, 4, 4</td>
<td>0.898214</td>
<td>0.115436</td>
<td>0.898214</td>
<td>0.115436</td>
<td>0.898214</td>
<td>0.115436</td>
<td>0.898214</td>
<td>0.115436</td>
</tr>
<tr>
<td>0.68, 0.71, 0.76, 0.81</td>
<td>2, 0, 1, 5</td>
<td>0.906714</td>
<td>0.086586</td>
<td>0.906714</td>
<td>0.086586</td>
<td>0.906714</td>
<td>0.086586</td>
<td>0.906714</td>
<td>0.086586</td>
</tr>
<tr>
<td>0.82, 0.835, 0.85, 0.865</td>
<td>9, 7, 3, 5</td>
<td>0.619643</td>
<td>0.258261</td>
<td>0.619643</td>
<td>0.258261</td>
<td>0.619643</td>
<td>0.258261</td>
<td>0.619643</td>
<td>0.258261</td>
</tr>
</tbody>
</table>

Table 9: Results of 10-fold cross-validation for binary threshold scheme using Source 2 and 3 features.

<table>
<thead>
<tr>
<th>Grade Threshold</th>
<th>Count Pass/Fail</th>
<th>DT Acc.</th>
<th>DT Std Dev.</th>
<th>RF Acc.</th>
<th>RF Std Dev.</th>
<th>SVM Acc.</th>
<th>SVM Std Dev.</th>
<th>MLP Acc.</th>
<th>MLP Std Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.700</td>
<td>77/2</td>
<td>0.973214</td>
<td>0.056626</td>
<td>0.973214</td>
<td>0.056626</td>
<td>1.0</td>
<td>0.0</td>
<td>0.973214</td>
<td>0.056626</td>
</tr>
<tr>
<td>0.750</td>
<td>77/2</td>
<td>0.973214</td>
<td>0.056626</td>
<td>0.987500</td>
<td>0.039528</td>
<td>1.0</td>
<td>0.0</td>
<td>0.973214</td>
<td>0.056626</td>
</tr>
<tr>
<td>0.800</td>
<td>71/8</td>
<td>1.000000</td>
<td>0.000000</td>
<td>0.987500</td>
<td>0.039528</td>
<td>1.0</td>
<td>0.0</td>
<td>0.973214</td>
<td>0.056626</td>
</tr>
<tr>
<td>0.725</td>
<td>69/10</td>
<td>1.000000</td>
<td>0.000000</td>
<td>1.000000</td>
<td>0.000000</td>
<td>1.0</td>
<td>0.0</td>
<td>0.973214</td>
<td>0.056626</td>
</tr>
<tr>
<td>0.750</td>
<td>65/18</td>
<td>1.000000</td>
<td>0.000000</td>
<td>1.000000</td>
<td>0.000000</td>
<td>1.0</td>
<td>0.0</td>
<td>0.973214</td>
<td>0.056626</td>
</tr>
</tbody>
</table>

Table 10: Results of 10-fold cross-validation for partition threshold scheme using Source 2 and 3 features.

<table>
<thead>
<tr>
<th>Partition Set</th>
<th>Count per Partition</th>
<th>DT Acc.</th>
<th>DT Std Dev.</th>
<th>RF Acc.</th>
<th>RF Std Dev.</th>
<th>SVM Acc.</th>
<th>SVM Std Dev.</th>
<th>MLP Acc.</th>
<th>MLP Std Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7, 0.75, 0.775, 0.8</td>
<td>2, 4, 4, 4</td>
<td>0.948214</td>
<td>0.089266</td>
<td>0.948214</td>
<td>0.089266</td>
<td>0.948214</td>
<td>0.089266</td>
<td>0.948214</td>
<td>0.089266</td>
</tr>
<tr>
<td>0.73, 0.775, 0.8, 0.83</td>
<td>2, 4, 4, 4</td>
<td>0.898214</td>
<td>0.115436</td>
<td>0.898214</td>
<td>0.115436</td>
<td>0.898214</td>
<td>0.115436</td>
<td>0.898214</td>
<td>0.115436</td>
</tr>
<tr>
<td>0.68, 0.71, 0.76, 0.81</td>
<td>2, 0, 1, 5</td>
<td>0.906714</td>
<td>0.086586</td>
<td>0.906714</td>
<td>0.086586</td>
<td>0.906714</td>
<td>0.086586</td>
<td>0.906714</td>
<td>0.086586</td>
</tr>
<tr>
<td>0.82, 0.835, 0.85, 0.865</td>
<td>9, 7, 3, 5</td>
<td>0.619643</td>
<td>0.258261</td>
<td>0.619643</td>
<td>0.258261</td>
<td>0.619643</td>
<td>0.258261</td>
<td>0.619643</td>
<td>0.258261</td>
</tr>
</tbody>
</table>

Table 11: Results of 10-fold cross-validation for partition threshold scheme using GPA, major, num_intern, and Source 2 and 3 features.

<table>
<thead>
<tr>
<th>Partition Set</th>
<th>Count per Partition</th>
<th>DT Acc.</th>
<th>DT Std Dev.</th>
<th>RF Acc.</th>
<th>RF Std Dev.</th>
<th>SVM Acc.</th>
<th>SVM Std Dev.</th>
<th>MLP Acc.</th>
<th>MLP Std Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7, 0.75, 0.775, 0.8</td>
<td>2, 4, 4, 4</td>
<td>0.948214</td>
<td>0.089266</td>
<td>0.948214</td>
<td>0.089266</td>
<td>0.948214</td>
<td>0.089266</td>
<td>0.948214</td>
<td>0.089266</td>
</tr>
<tr>
<td>0.73, 0.775, 0.8, 0.83</td>
<td>2, 4, 4, 4</td>
<td>0.898214</td>
<td>0.115436</td>
<td>0.898214</td>
<td>0.115436</td>
<td>0.898214</td>
<td>0.115436</td>
<td>0.898214</td>
<td>0.115436</td>
</tr>
<tr>
<td>0.68, 0.71, 0.76, 0.81</td>
<td>2, 0, 1, 5</td>
<td>0.906714</td>
<td>0.086586</td>
<td>0.906714</td>
<td>0.086586</td>
<td>0.906714</td>
<td>0.086586</td>
<td>0.906714</td>
<td>0.086586</td>
</tr>
<tr>
<td>0.82, 0.835, 0.85, 0.865</td>
<td>9, 7, 3, 5</td>
<td>0.619643</td>
<td>0.258261</td>
<td>0.619643</td>
<td>0.258261</td>
<td>0.619643</td>
<td>0.258261</td>
<td>0.619643</td>
<td>0.258261</td>
</tr>
</tbody>
</table>
3.3 Clustering Efficacy
3.3.1 Visualization with Dendrograms

The dendrograms shown in Figures 4 through 11 visualize the effects that feature selection and weighting (uniform and non-uniform alike) have on the single-linkage Euclidean distance-based clustering algorithm. Each decimal value in the weights list seen in the title of each figure represents the weight of a particular subset of features. The breakdown of which features are represented by which weights is described in the following respective order: demographics and gender identity, major, number of internships, GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data. Figures 6 through 11 only list four decimal values in their titles; these correspond to GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data, respectively. The preceding features are not included i.e. have zero weight. A given dendrogram’s y-axis indicates the distance between the merged clusters while the x-axis shows each record’s index in the data set.

Weights: [1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]

Figure 4: Clustering with uniform weighting and all features present.
When considered together, Figures 4 and 5 are able to demonstrate the effects of feature selection on the distance between records. Across all the records (seen along the x-axis), there is a noticeable decrease in the distances between records after the demographics and gender identity features are removed; this change in distance is also showcased by the change in color, where the color change threshold was set to be when the distance reached 2.75 (as reflected in the y-axis). Both figures demonstrate a recurring phenomenon with what could be considered more "weakly" clustered records, and this is the sequential adding of individual records to a cluster rather than combining whole clusters together. For the purposes of this study, this phenomenon can henceforth be referred to as "chain-link clustering".

Figure 4 shows this in two particular regions: the second quarter of records from the left and the rightmost third of the graph (the latter to a lesser degree due to these records mostly just being outliers; consider the distances when clustering). Conversely, only the rightmost portion of Figure 5 demonstrates chain-link clustering. Note the
"stronger" clustering in the leftmost portion of Figure 5; rather than individual records being sequentially added to a larger cluster, smaller clusters that form from those records later merge into a single larger cluster.

Figure 6, which completely removes all Source 1 features (except for GPA) shows even stronger clustering than seen in Figure 5, which kept the "major" feature. It too suffers from chain-link clustering, but this is, again, only observed in the rightmost third of the dendrogram. Distances are once again reduced, but only marginally; the significance of this comparison lies in how many smaller clusters are created before larger ones are formed through their merging. In this manner, Figure 6 thus demonstrates a clear improvement over Figures 4 and 5 in that the clustering can be considered stronger.

Figure 6: Uniformly weighted clustering without demographics, gender identity, major, or number of internships. Only GPA, weekly survey averages, number of weekly labs attended are present, and pre-course survey data.

However, the same cannot be said for Figure 7 which completely removes the Source 3 features i.e. pre-course survey data. In comparing with Figure 6, a
massive reduction in distances between merged clusters and records is observed; this decrease is so large the y-axis scale (which corresponds to the distance between two merged clusters/records) in turn reduces from a maximum of 7 to a maximum of 3.

Figure 7: Uniformly weighted clustering without demographics, gender identity, major, number of internships, or pre-course survey data. Only GPA, weekly survey averages, and number of weekly labs attended are present.

It should be noted, though, that chain-link clustering is decidedly less prevalent in Figure 7 than in Figure 6; records are being grouped together rather than being added sequentially to larger clusters, especially in the rightmost third of the graph. Where many of these records were, in effect, outliers to all other records, the removal of the Source 3 features has allowed them to initially form smaller clusters before joining the larger collective.
Another angle from which this question can be tackled is by switching which features are selected. Figure 8 reinstates the Source 3 features that were absent in Figure 7 and instead removes the feature indicating the number of weekly labs attended by zeroing out its affiliated weight. In doing so, Figure 8 bears a much greater similarity to Figure 6 in that the scale of distances matches more closely; while there is an overall reduction in distance, it is nowhere near as large as the one observed in Figure 7. It should be expected, then, that removing both the pre-course survey features as well as the number of weekly labs attended renders a dendrogram with even smaller distances and a greater number of smaller clusters initially formed; this is Figure 9, which has practically negligible chain-link clustering.
In showcasing the changes between these figures, it has become extremely clear that the amount of chain-link clustering present is tied inherently to the distances. What Figure 9 details, though, by only utilizing very few features and maintaining a particularly small distance scale, is that limiting the number of features in order to obtain that (if only visually) "stronger" clustering may not generate the models that could actually be considered worthwhile. This observation is investigated further in Section 3.3.2. For the time being, however, it can be assumed (based on the findings of Section 3.1) that a more optimal clustering parameter vector would not minimize the distances in a blanket manner, but would rather assign greater weight to the more significant features and either decrease or zero out the weights of the less important ones.

In keeping with this assumption, Figures 10 and 11 are generated by adjusting the weights to better match the Section 3.1 findings. Figure 10 presents an initial step...
building off of Figure 6 by only increasing the weight of the GPA feature to 1.2 and keeping the Source 2 and 3 features of normal weight (i.e. 1.0).

Figure 10: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. GPA is weighted more to reflect its greater significance according to the feature rankings detailed in Section 3.1.

Figure 11 iterates further by keeping the weight of the GPA feature as 1.2 but adding to the weekly survey features’ weight, making it 1.1, with the Source 3 features’ weight kept at 1.0. This parameter configuration is designed to mimic the Section 3.1 findings more closely than any of the preceding dendrograms. Emphasizing the GPA, weekly survey averages, and pre-course survey features (in descending order of significance) causes two phenomena: first, there is a general increase in distances across the dendrogram, and second, there are a few more more smaller clusters forming towards the right end of Figure 11 than Figure 10. This is in spite of both the general distance increase and the fact that both figures have roughly the same amount of chain-link clustering. That the distance gain does not induce "weaker" clustering
is particularly notable given the pattern ascertained across other dendrograms runs contradictory to it.

Figure 11: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1.

3.3.2 Individual Records Case Study

Visualizing the clustering patterns with dendrograms is not quite enough to truly understand the effects that the feature selection and weighting have on how the records are actually clustered. Because the font along the x-axis (detailing the records’ actual indices) is too small to read in the dendrograms as they are presented above, the figures do not clearly communicate which records are clustered together. So, in order to more effectively grasp the effects of the weighting and choice of features, a number of specific records will be followed across the dendrograms utilizing different weights. This will be done for six "positive" and seven "negative" records. Recall, as described near the end of Chapter 2, that the grade threshold which dictates this positive/negative state is 0.8. With regard to the visualization itself, both the original
dendrograms as well as specifically zoomed-in figures are provided, all annotated to highlight the records’ indices so their relative positions can be more closely observed.

### 3.3.2.1 Positive Cases

![Figure 12](image)

Figure 12: Clustering with uniform weighting and all features present. Positive cases are highlighted.

With this first weighting seen in Figures 12 and 13, all features are included and weighted uniformly. It can already be observed that the records are mostly scattered among the other records, but take note of their general placement in the dendrogram; the records are firmly in the leftmost two-thirds of the Figure 12. Coupled with their being merged with low-to-median distances (according to the y-axis), this positioning indicates that they are decidedly not outliers like the ones seen in rightmost portion of Figure 12. These outliers are omitted from Figure 13, which presents a more zoomed-in image that allows an easier-to-read observation.

As seen in Figure 13, records 7 and 55 are shown to be closer to each other than to any of other tracked records. Compare this to Figures 14 and 15, where all the records are clustered into a small region, but the aforementioned records are now apart.
Figure 13: Clustering with uniform weighting and all features present. Positive cases are highlighted. Zoomed in.

Figure 14: Uniformly weighted clustering without demographics and gender identity. Positive cases are highlighted.
Figures 14 and 15 showcase a moderate general reduction in distance between clustered records, but the most significant observation in comparison to Figures 12 and 13 is the rearrangement of records. This ultimately will provide insight into the effects of feature selection and the importance of proper weighting. Note, however, that this rearrangement is still contained to just a small number of records. As was observed in Figure 5, only the leftmost third of the dendrogram demonstrates "stronger" clustering. Later figures will show more of this "stronger" clustering, but emphasis will be placed on the individual records in order to gauge the clustering trials’ actual quality.

Compared to Figures 14 and 15, the features being removed in Figures 16 and 17 are not very significant according to the results of Section 3.1. It is therefore not surprising that, in addition to the clustering looking "stronger", the records are still grouped together in the same fashion. Later figures such as 18, 20, and 22 remove different features of varying importance. The consequences of these changes range
from minimal (some separation) to drastic (complete detachment).

Figure 16: Uniformly weighted clustering without demographics, gender identity, major, or number of internships. Only GPA, weekly survey averages, number of weekly labs attended are present, and pre-course survey data. Positive cases are highlighted.

Figure 17: Uniformly weighted clustering without demographics, gender identity, major, or number of internships. Only GPA, weekly survey averages, number of weekly labs attended are present, and pre-course survey data. Positive cases are highlighted. Zoomed in.
Figures 18 and 19 demonstrate this drastic consequence. The removal of the Source 3 features causes the cohesion between the records (observed since Figure 14) to almost totally disintegrate. Records 7 and 47 are now in separate clusters, 55 is now closer to 15 rather than 34, and 31 is completely removed from the rest.

Figure 18: Uniformly weighted clustering without demographics, gender identity, major, number of internships, or pre-course survey data. Only GPA, weekly survey averages, and number of weekly labs attended are present. Positive cases are highlighted.

In contrast, Figures 20 and 21 are decidedly more subtle in their changes from Figure 16 than Figure 18 is. Record 31 is still removed from the other records, but by a much smaller degree. Furthermore, the remaining records maintain their cohesion and order completely.
Figure 19: Uniformly weighted clustering without demographics, gender identity, major, number of internships, or pre-course survey data. Only GPA, weekly survey averages, and number of weekly labs attended are present. Positive cases are highlighted. Zoomed in.

Weights: [1.0, 1.0, 1.0, 0]

Figure 20: Clustering with uniform weighting without demographics, gender identity, major, number of internships, or number of weekly labs attended. Only GPA and weekly survey averages are present. Positive cases are highlighted.

Weights: [1.0, 1.0, 0, 1.0]
Figure 21: Clustering with uniform weighting without demographics, gender identity, major, number of internships, or number of weekly labs attended. Only GPA, weekly survey averages, and pre-course survey data. Only GPA and weekly survey averages are present. Positive cases are highlighted. Zoomed in.

Figures 22 and 23 carry a number of interesting observations. First and foremost is that the records are still near each other in spite of the Source 3 data being missing. Recall that in Figures 18 and 19, the records were spread widely across the left two-thirds of the dendrogram; the difference with Figures 22 and 23 is that the number of weekly labs attended is also zeroed out. Aside from the feature vector, the records’ order itself has been greatly rearranged. Record 47 (rather than 31) is now alone in a completely separate cluster, 7 and 47 are at opposing ends of the record sequence, and 55 is further apart from 15 and 34 which, in addition to 31, are only connected to 7 and 55 by way of chain-link clustering.
Figure 22: Uniformly weighted clustering without demographics, gender identity, major, number of internships, number of weekly labs attended, or pre-course survey data. Only GPA and weekly survey averages are present. Positive cases are highlighted.

Figure 23: Uniformly weighted clustering without demographics, gender identity, major, number of internships, number of weekly labs attended, or pre-course survey data. Only GPA and weekly survey averages are present. Positive cases are highlighted. Zoomed in.
Figure 24: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. GPA is weighted more to reflect its greater significance according to the feature rankings detailed in Section 3.1. Positive cases are highlighted.

With Figures 24 and 25, the GPA is weighted more highly with a 1.2 coefficient whereas the Source 2 and 3 features are all restored and maintain a 1.0 weight. Also restored is the records’ order and clustering pattern, with 7 and 47 forming an initial cluster, 34 and 55 doing the same, 15 merging with the latter, all five records coming together, and then 31 finally being added.

Figures 26 and 27 match the original feature rankings even more closely and show no change with the six records being tracked. Given that incrementally altering the weights to align with the feature rankings has shown continual improvements with these six records, it stands to reason that while these particular records did not change in this iteration, other records may ultimately cluster better as a result. Even if there are no records in this data set that follow this behavior, recall that the set utilized in this study is rather small. Having more data would help further corroborate this point in the future.
Figure 25: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. GPA is weighted more to reflect its greater significance according to the feature rankings detailed in Section 3.1. Positive cases are highlighted. Zoomed in.

Figure 26: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. Positive cases are highlighted.
3.3.2.2 Negative Cases

Throughout most of the following figures, the seven "negative" cases are observed to have few commonalities to the point that they are generally not clustered together. Figures 28 and 29 are exceptional in that, while the records are seen as not being very close to each other, their positions are not extremely widespread. This is in stark contrast to dendrograms with weighting configurations that more closely resemble the results from Section 3.1. Also note that, compared to the records followed in Figure 12, these cases are much more condensed around the mid-section of the graph where chain-link clustering is more prevalent. This is ultimately just further evidence of the records’ dissimilarity.
Figure 28: Clustering with uniform weighting and all features present. Negative cases are highlighted.

Figure 29: Clustering with uniform weighting and all features present. Negative cases are highlighted. Zoomed in.
Similar to Figures 18 and 19, Figures 30 and 31 show the disintegration of the records’ cohesion. The difference is that this occurred as a result of removing less important features, rather than features that were known to be significant.

Figure 30: Uniformly weighted clustering without demographics and gender identity. Negative cases are highlighted.

Figure 31: Uniformly weighted clustering without demographics and gender identity. Negative cases are highlighted. Zoomed in.
Figures 32 and 33 demonstrate a marginal recollection of the records, but, like Figures 28 and 29 they are clearly the exception to the norm. The trend of lacking cohesion among the "negative" cases continues with the remaining figures.

Figure 32: Uniformly weighted clustering without demographics, gender identity, major, or number of internships. Only GPA, weekly survey averages, number of weekly labs attended are present, and pre-course survey data. Negative cases are highlighted.
Figure 33: Uniformly weighted clustering without demographics, gender identity, major, or number of internships. Only GPA, weekly survey averages, number of weekly labs attended are present, and pre-course survey data. Negative cases are highlighted. Zoomed in.

Figure 34: Uniformly weighted clustering without demographics, gender identity, major, number of internships, or pre-course survey data. Only GPA, weekly survey averages, and number of weekly labs attended are present. Negative cases are highlighted.
Figure 35: Uniformly weighted clustering without demographics, gender identity, major, number of internships, or pre-course survey data. Only GPA, weekly survey averages, and number of weekly labs attended are present. Negative cases are highlighted. Zoomed in.

Figure 36: Clustering with uniform weighting without demographics, gender identity, major, number of internships, or number of weekly labs attended. Only GPA, weekly survey averages, and pre-course survey data Only GPA and weekly survey averages are present. Negative cases are highlighted.
Figure 37: Clustering with uniform weighting without demographics, gender identity, major, number of internships, or number of weekly labs attended. Only GPA, weekly survey averages, and pre-course survey data. Only GPA and weekly survey averages are present. Negative cases are highlighted. Zoomed in.

Weights: [1.0, 1.0, 0, 1.0]

Figure 38: Uniformly weighted clustering without demographics, gender identity, major, number of internships, number of weekly labs attended, or pre-course survey data. Only GPA and weekly survey averages are present. Negative cases are highlighted.

Weights: [1.0, 1.0, 0, 0]
Figure 39: Uniformly weighted clustering without demographics, gender identity, major, number of internships, number of weekly labs attended, or pre-course survey data. Only GPA and weekly survey averages are present. Negative cases are highlighted. Zoomed in.

Figure 40: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. GPA is weighted more to reflect its greater significance according to the feature rankings detailed in Section 3.1. Negative cases are highlighted.
Figure 41: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. GPA is weighted more to reflect its greater significance according to the feature rankings detailed in Section 3.1. Negative cases are highlighted. Zoomed in.

Figure 42: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. Negative cases are highlighted.
Figure 43: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. Negative cases are highlighted. Zoomed in.

Figures 42 and 43, in spite of their weight configuration’s effectiveness in grouping the "positive" cases, ultimately do little to establish cohesion between the "negative" cases. There is some connectivity between records 27, 66, and 67, but even with those records, the general theme is the same: the "negative" cases do not cluster together directly. Each of them are outliers from other "negative" records.
3.3.3 Confusion Matrices

In this scheme, the objective is to understand how modifications to the feature vector (either by selection or weighting) alter the numerical metrics of true/false positive/negative counts. Results for a total of three different grade thresholds across all eight previously considered weight configurations are reported. Of the records in this data set, twenty-three have grades below 0.87, eight below 0.8, and five below 0.78. Recall from Chapter 2 the idea of a cluster size threshold being employed in order to determine the point at which the clustering trial is to be stopped prematurely so the confusion matrix can be calculated. The cluster size data reported in this section is not the threshold parameter itself, but rather, as explained in that chapter, the actual size of the cluster when termination happens. The cluster size thresholds used to generate the figures and table rows were 26 and 39 (roughly one-third and one-half the size of the data set, respectively). Each weight configuration thus appears in a pairwise manner, one for the first cluster that exceeds a size of 26 records, and the second for the first cluster that exceeds size 39. The main exception to this is the uniform weighting (1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0) as the first cluster that exceeds size 26 is a cluster of size 40, which is also greater than 39. Because both thresholds generate the same information, only one instance is included. The other exceptional case is the weighting configuration of (1.2, 1.1, 1.0, 1.0) when the grade threshold is 0.8. Because this weighting most closely resembles the feature significance ranking of Section 3.1, its experimentation has been expanded upon by testing many more cluster size thresholds, listed as follows: 10, 15, 20, 26, 29, 30, 33, 39, 45, 50, 55, 60, 65, 70. Across these fourteen threshold parameter values, thirteen unique confusion matrices were generated.
Table 12: Confusion matrix results.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Threshold</th>
<th>Features</th>
<th>TN Rate/Count</th>
<th>FN Rate/Count</th>
<th>TP Rate/Count</th>
<th>FP Rate/Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.78</td>
<td>0.1,0.1,0.1,0.1,0.1,0.1,0.1</td>
<td>40</td>
<td>0.0256/1</td>
<td>0.9744/38</td>
<td>0.9000/36</td>
<td>0.1000/4</td>
</tr>
<tr>
<td>0.78</td>
<td>0.1,0.1,0.1,0.1,0.1,0.1,0.1</td>
<td>26</td>
<td>0.0755/4</td>
<td>0.9245/49</td>
<td>0.9015/25</td>
<td>0.0385/1</td>
</tr>
<tr>
<td>0.78</td>
<td>0.1,0.1,0.1,0.1,0.1,0.1,0.1</td>
<td>39</td>
<td>0.0750/3</td>
<td>0.9250/37</td>
<td>0.9487/37</td>
<td>0.0513/2</td>
</tr>
<tr>
<td>0.78</td>
<td>0.1,0.1,0.1,0.1,0.1,0.1,0.1</td>
<td>28</td>
<td>0.0784/4</td>
<td>0.9216/47</td>
<td>0.9643/27</td>
<td>0.0357/1</td>
</tr>
<tr>
<td>0.78</td>
<td>0.1,0.1,0.1,0.1,0.1,0.1,0.1</td>
<td>39</td>
<td>0.1000/4</td>
<td>0.9000/36</td>
<td>0.9744/38</td>
<td>0.0256/1</td>
</tr>
<tr>
<td>0.78</td>
<td>0.1,0.1,0.1,0.1,0.1,0.1,0.1</td>
<td>27</td>
<td>0.0577/3</td>
<td>0.9423/49</td>
<td>0.9259/25</td>
<td>0.0741/2</td>
</tr>
<tr>
<td>0.78</td>
<td>0.1,0.1,0.1,0.1,0.1,0.1,0.1</td>
<td>46</td>
<td>0.0606/2</td>
<td>0.9394/31</td>
<td>0.9348/43</td>
<td>0.0652/3</td>
</tr>
<tr>
<td>0.78</td>
<td>0.1,0.1,0.1,0.1,0.1,0.1,0.1</td>
<td>26</td>
<td>0.0755/4</td>
<td>0.9245/49</td>
<td>0.9015/25</td>
<td>0.0385/1</td>
</tr>
<tr>
<td>0.78</td>
<td>0.1,0.1,0.1,0.1,0.1,0.1,0.1</td>
<td>45</td>
<td>0.1176/4</td>
<td>0.8824/30</td>
<td>0.9778/44</td>
<td>0.0222/1</td>
</tr>
<tr>
<td>0.78</td>
<td>0.1,0.1,0.1,0.1,0.1,0.1,0.1</td>
<td>34</td>
<td>0.0667/3</td>
<td>0.9333/42</td>
<td>0.9412/32</td>
<td>0.0588/2</td>
</tr>
<tr>
<td>0.78</td>
<td>0.1,0.1,0.1,0.1,0.1,0.1,0.1</td>
<td>50</td>
<td>0.1034/3</td>
<td>0.8966/26</td>
<td>0.9600/48</td>
<td>0.0409/2</td>
</tr>
<tr>
<td>0.78</td>
<td>1.2,1.0,1.0,1.0,1.0,1.0,1.0</td>
<td>26</td>
<td>0.0755/4</td>
<td>0.9245/49</td>
<td>0.9015/25</td>
<td>0.0385/1</td>
</tr>
<tr>
<td>0.78</td>
<td>1.2,1.0,1.0,1.0,1.0,1.0,1.0</td>
<td>39</td>
<td>0.1000/4</td>
<td>0.9000/36</td>
<td>0.9744/38</td>
<td>0.0256/1</td>
</tr>
<tr>
<td>0.78</td>
<td>1.2,1.1,1.0,1.0,1.0,1.0,1.0</td>
<td>26</td>
<td>0.0755/4</td>
<td>0.9245/49</td>
<td>0.9015/25</td>
<td>0.0385/1</td>
</tr>
<tr>
<td>0.78</td>
<td>1.2,1.1,1.0,1.0,1.0,1.0,1.0</td>
<td>39</td>
<td>0.1000/4</td>
<td>0.9000/36</td>
<td>0.9744/38</td>
<td>0.0256/1</td>
</tr>
</tbody>
</table>
Every row in Table 12 that has a grade threshold of 0.8 is emphasized visually by having the same information presented in Figure form. Figures 44 to 69 each show a confusion matrix display as generated by the Python "sklearn" library’s "metrics" module based on the grade threshold, weights, and cluster size threshold parameters. The former two parameters and largest cluster’s size are all shown directly in each figure, as are the true/false positive/negative counts and derived rates. What is easily observable (and logical, based on the scheme’s design), is that the cluster size affects how many records are labeled "negative" and how many are "positive". Conversely, the weight configuration determines the arrangement of the records and thus affects the true/false nature of the aforementioned labelling.
Figure 44: Confusion matrix of clustering with uniform weighting and all features present. "Positive" cluster size is 40.
In Chapter 2, it was explained that high true negative and positive rates are ideal for these weight configurations. Considering the uniformly-weighted nature of Figure 44, it is sensible to let the results of this specific configuration act as a baseline for comparison. The true negative and true positive rates in Figure 44 are 0.0769 and 0.8750, respectively. Recall that, considering these figures all utilize the 0.8 grade threshold, the number of actual negatives in the data set is eight. Of those eight records, according to the figure, only three are correctly predicted to be "negative", meaning the actual accuracy of the negative prediction is three out of eight, or 37.5 percent.

Figure 45: Uniformly weighted clustering without demographics and gender identity. "Positive" cluster size is 26.
With this percentage as the starting point, it will be interesting to note how the true negative rate increases in later figures i.e. as the weight configuration is adjusted to match the Section 3.1 results. The cluster size will also need to be factored in, as it carries a significant effect in how many records are predicted to be "negative".

Figure 46: Uniformly weighted clustering without demographics and gender identity. "Positive" cluster size is 39

Figures 45 and 46 showcase a substantial increase in both the true negative and true positive rates. Furthermore, while the true positive rate remains constant between these two figures, the true negative rate is actually higher in Figure 46.
Figure 47: Uniformly weighted clustering without demographics, gender identity, major, or number of internships. Only GPA, weekly survey averages, number of weekly labs attended are present, and pre-course survey data. "Positive" cluster size is 28.

Figures 47 and 48 demonstrate an even greater increase in both rates. Note however, that while the true negative rate increases in Figure 48 compared to Figure 47, the true positive rate actually decreases. The only different parameter between these two figures is the cluster size, so the placement of the "negative" cases is a clearly important, yet notably (in a practical context) uncontrollable factor that needs to be considered.
Figure 48: Uniformly weighted clustering without demographics, gender identity, major, or number of internships. Only GPA, weekly survey averages, number of weekly labs attended are present, and pre-course survey data. "Positive" cluster size is 39.

Given the complete breakdown of any cohesion between individual records that was observed in Figures 18 and 34 coupled with the fact of "negative" record placement affecting the confusion matrix results, it would stand to reason that the removal of the pre-course survey data likely means the model cannot accurately label the "negative" cases, especially not with a larger cluster size parameter. This prediction proves true, as Figures 49 and 50 demonstrate a substantial decrease in both the true negative and true positive rates from the findings of Figures 47 and 48.
Take note of the count of predicted negatives in these two figures. Figure 49 has a much larger net being cast, with a total of 52 records being predicted negative, six of which are true. Figure 50 has only 33 negative predictions and it effectively captures four of the true "negative" cases. The difference in ratio contributes to Figure 50 having the higher true negative rate. However, it also has a lower true positive rate, a result of having two more (twice as many) false positives and a total number of positive predictions less than twice that of Figure 49.

![Figure 49: Uniformly weighted clustering without demographics, gender identity, major, number of internships, or pre-course survey data. Only GPA, weekly survey averages, and number of weekly labs attended are present. "Positive" cluster size is 27.](image-url)
Figure 50: Uniformly weighted clustering without demographics, gender identity, major, number of internships, or pre-course survey data. Only GPA, weekly survey averages, and number of weekly labs attended are present. "Positive" cluster size is 46.
Figure 51, which restores the pre-course survey data but removes the weekly lab attendance variable, has an even larger group labeled "negatives" than Figure 49. However, due to the weight configuration’s rearrangement of the record placement, it captures even fewer true negatives. Its true negative and true positive rates are lower as a result.

Figure 51: Clustering with uniform weighting without demographics, gender identity, major, number of internships, or number of weekly labs attended. Only GPA, weekly survey averages, and pre-course survey data Only GPA and weekly survey averages are present. “Positive” cluster size is 26.

Figure 52 fares a bit better with higher rates, but its larger "positive" cluster size
causes it to lose two of the "negative" cases.

Figure 52: Clustering with uniform weighting without demographics, gender identity, major, number of internships, or number of weekly labs attended. Only GPA, weekly survey averages, and pre-course survey data Only GPA and weekly survey averages are present. "Positive" cluster size is 45.
Having reached Figures 53 and 54, the trend is starting to become very clear; removing significant features generally decreases the accuracy of the clustering process, regardless of which size is being used. Even if the cluster size is small, there is no guarantee that the students who are at-risk will actually be labeled as "negative". And if the cluster size is large, most of the "negative" cases are labeled as "positive" as part of a massive cluster, making them tedious and difficult for staff to root out.

Figure 53: Uniformly weighted clustering without demographics, gender identity, major, number of internships, number of weekly labs attended, or pre-course survey data. Only GPA and weekly survey averages are present. "Positive" cluster size is 34.
Figure 54: Uniformly weighted clustering without demographics, gender identity, major, number of internships, number of weekly labs attended, or pre-course survey data. Only GPA and weekly survey averages are present. "Positive" cluster size is 50.
With Figure 55, progress is made once again. It has the highest true positive rate yet seen, the second highest true negative rate (behind only Figure 48), and captures all but one of the "negative" cases. Furthermore, it accomplishes all of this with a smaller cluster size of only 26. This leap in effectiveness compared to previous figures should be unsurprising considering the feature vector involved. All the significant features are included, and the GPA is more heavily weighted to further emphasize its importance amongst the rankings seen in Section 3.1.

Figure 55: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. GPA is weighted more to reflect its greater significance according to the feature rankings detailed in Section 3.1. "Positive" cluster size is 26.

![Figure 55](image-url)
Figure 56: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. GPA is weighted more to reflect its greater significance according to the feature rankings detailed in Section 3.1. "Positive" cluster size is 39.

Figure 56 maximizes the cluster size parameter, causing a general decrease in negative predictions while only losing one true "negative" record for total of six out of eight captured. Subsequently, its true negative rate is higher than even that of Figure 55; at 0.15, the true negative rate is the highest seen yet. Both the number of captured "negatives" as well as the true negative rate have far surpassed what was initially seen in Figure 44. This does, however, again come at the cost of a lower true positive rate. The pattern of figure performance seems to indicate that a cluster size
of 39 is about as close to ideal as can be found with this data set. A mostly-equal split between negative and positive predictions has been generating some of the best rates. In order to better corroborate this theory, a more thorough breakdown of the (1.2, 1.1, 1.0, 1.0) configuration, which even more closely matches the Section 3.1 findings, is in order, starting with Figure 57. As expected, labeling a very large portion of the data set (66 out of 79 records) as being "negative" causes the correct prediction of most of the truly "negative" cases, but leads to a severe decrease in the true negative rate.

Figure 57: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 13.
Figure 58: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 15.

Stepping incrementally forward in the cluster size decreases the number of total "negative" predictions. Between Figures 58, 59, 60, and so on, a steady increase in the true negative and positive rates can be observed.
Figure 59: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 22.
Figure 60: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 26.
Figure 61 is intriguing because of its notable decrease in rates. Take note of how, compared to the previous figures employing the (1.2, 1.1, 1.0, 1.0) weight configuration, the number of true "negatives" (i.e. the top two quadrants of the figure) changes from seven and one to six and two. This continues up to and including Figure 63, which has identical rates and counts to Figure 56. Recall that the (1.2, 1.1, 1.0, 1.0) weight configuration, while having larger distances, showed better clustering, indicating a higher quality parameter.

Figure 61: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 32.
Figure 62: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 33.
Figure 63: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 39.
Figure 64 showcases a massive decrease in quality in that only three out of eight negative records are accurately predicted. Going forward, a downhill trend in accurate prediction of "negative" records is observed, and while the increase in cluster size means there are fewer records labeled "negative", thus occasionally showing increases in the true negative and positive rates, the lack of accurate labeling means these higher cluster sizes are not anywhere close to ideal for this weight configuration.

Figure 64: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 46.
Figure 65: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 50.
Figure 66: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 56.
Figure 67: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 61.
Figure 68: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 65.
Figure 69: Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 70.
CHAPTER 4

Discussion

4.1 Making a Statement with ML Classifiers

The bar graphs and model accuracy/standard deviation tables above make it extremely clear that the survey responses in Sources 2 and 3 and academic history detailed by GPA are of great significance. This is, as expected, in contrast to the students’ gender identity and especially demographics. Ethnicity and racial profile are scientifically known to have absolutely no effect on an individual’s ability to learn or perform. With these results, it is ultimately proven that any correlation between such identifiers and academic performance cannot be found by directly comparing the two concepts in a vacuum; rather, it is necessary to understand the presence of the forces, circumstances, and situations completely external to the students, their potential, and their capabilities that greatly affect them and their ability to succeed. This is further discussed by many of the articles described in the related study section and, again, a key purpose of UNVS 101 is to try to understand and address those exact problems. What makes the survey features so much more meaningful to the models than the demographics is that the survey responses reflect entirely on the problems plaguing underserved/disadvantaged students. Those demographics are effectively just numbers; they have, as is scientifically proven, nothing inherently to do with performance, potential, learning ability, or job readiness, and it would therefore be foolish to consider them as being indicative of any of these attributes. This is why Table 11 shows the results of models that deliberately exclude all of the demographics and gender identity features. The fact that it returns a slightly higher accuracy than Table 10 which itself is higher than Table 6 shows the outright harm that including these personal identity indicators can do to a system that is designed to determine an individual’s status as successful or failing. That is not to say that
those features can never have a place in models such as these; the restriction is that they should not be included without additional features that are explicitly about those external problems and circumstances as well as their effects. Given the known correlation between educational attainment and family income, for example, it would stand to reason that at least some of those additional columns would need to be primarily focused on the financial backgrounds of the students. In order to examine, classify, and/or group students accurately, a scientific approach that considers more features (i.e. employs a larger context) is necessary. Taking only one or a few of these features misses a majority of necessary information to build correct models.

4.2 The Inadequacy of PCA

All of that said, though, if Figures 1, 2, and 3 provide such proof of the Source 2 and 3 features’ importance, then why are the PCA results conflicting with those findings? Ultimately, this dissonance stems from PCA being misused in this study in two different ways. The first is that this particular ML technique is not meant to be used with non-continuous (or “discrete”) data, a description which exactly fits most of the features in this project’s data set. Recall Table 1, where the specific features are each listed alongside brief notes on their possible values. Consider the Source 1 features, where only the student’s GPA (excluding the target label “grade”) has an infinite number of possible values; GPA would qualify as being “continuous” where the remaining features, which have only a limited number of possible values, would be considered “discrete”. Source 2 is the exact opposite, where nearly all of its features are continuous with the exception of “num_labs_attended”, and Source 3 is filled entirely with discrete features. When considering all three sources, there is a clearly overwhelming emphasis on discrete variables rather than continuous ones, and
so PCA would be a very poor choice of technique to use with this data set. The second way in which PCA was misused is less specific to this data set. PCA is generally ineffective as a method of feature selection and, by extension, scoring and ranking. This ineffectiveness can be traced to the computational design of PCA, where the magnitude of a feature’s variance is the ultimate determinant of its significance. An example of this can be found in Table 2; consider how “can_make_career” is ranked very highly while GPA is near the bottom of the ranking. Recall that the former operates on a scale of 1 to 7 while the latter must be a decimal between 0 and 4, inclusive. Barring an extremely bizarre distribution of records for these two features, it is clear that “can_make_career” will have a higher variance than GPA based solely on the fact that its scale is of a wider range. A closer look at the formula for calculating variance, which involves squaring the differences between each record’s value and the set’s mean value, indicates that either the responses to “can_make_career” are not necessarily clustered around the mean, the GPA values are more clustered than the “can_make_career” responses, or both.

Counting the number of occurrences for each individual numerical response to “can_make_career” using Microsoft Excel finds that clustering is only moderate; out of the 79 rows, where the mean response is 5.67, one individual provided an answer of “2”, two responded with “3”, six with “4”, twenty-three with “5”, twenty-eight with “6”, and nineteen with “7”. For this variable to have a small variance, the responses would need to be specifically clustered within a very small range of values; because the responses are discrete and only have a handful of available answers, a small variance would necessitate the gathering of responses around a single primary value. Instead, there is a very wide distribution of responses around "5", "6", and "7", leading to a larger variance. It should also be noted that a distribution where responses are mostly gathered around a single value would be highly improbable given the nature
of the questions and the variety of students asked.
For example, a closer look at GPA can be taken. Most of the values in this column range from about 2.4 to 4.0 and are largely centered around their mean value of 3.23. Calculators would read this overall range as being around or lower than 1.6 whereas “can_make_career” has a range of at least 3 considering its highlighting the values "5", "6", and "7". Microsoft Excel’s "VAR" function states the variance for the GPA is only 0.164 while “can_make_career” has a variance of 1.172. This emphasizes the significant effect that a smaller range has in reducing the overall variance. In spite of the fact that “can_make_career” is somewhat clustered and GPA is not, the PCA results show GPA to be near the bottom of the table while “can_make_career” tops the chart alongside many of its fellow 7-scale discrete variables. What this ultimately concludes is that the only way that PCA could be a viable method of identifying feature significance is if the order of the features by significance perfectly matched their order by variance (the most important are the most varying, the least important are the least varying) which is an extremely unlikely occurrence. Not only was PCA a clearly poor choice of technique to use with this data set on account of its being designed for usage with continuous rather than discrete data, the reasons for which it was employed (feature ranking) was a complete departure from its intended purpose.

4.3 Confirmation with Clustering
4.3.1 Visualization: Exploring Visible Feature Significance

It was established at the start of Section 3.3.2 that visualization alone was not enough to provide indisputable proof of the GPA, weekly survey, and pre-course survey features’ significance, but it has still proven helpful in showcasing the effects made by changes in the features’ weighting. By combining the visualization power of clustering with the parameterization of feature selection and weighting, baseline observations
of how different weighting configurations affect the quality of the readout are made. The visualizations start with the uniform weight configuration and incrementally transition toward matching the feature significance ranking. As expected, massive shifts in model quality are continually observed as a direct result of changes in weight configuration. Compare, for example, Figures 4 and 5, where "stronger" clustering can be instantly observed just by removing a subset of less-significant features (in this case, demographics and gender identity). Recalling what is known and stated earlier in Section 4.1 about the scientific fallacies of focusing on these kinds of features without appropriate context, this observed positive change in removing them is not surprising. What the visualization does uniquely point out, however, is how much more similar the records become as the features’ weights adjust to match Section 3.1’s findings. Notice that the "stronger" clustering replaces the chain-link clustering patterns observed in many of the less-ideal weight configurations, especially as the distances between the records shrink. This is (reasonably) contradicted once the (1.2, 1.1, 1.0, 1.0) configuration is utilized in Figure 11, though. There is, as expected, more chain-link clustering with an overall greater weight than, say, in the (1.0, 1.0, 1.0, 1.0) configuration. However, it is likely that the weight configurations that more closely match Section 3.1’s findings not only generate better results that are readily observable but ensure that other records cluster more appropriately. This is not confirmed, but extremely probable. The next scheme allows for closer examination of those records that are clustered in this chain-link fashion, making it clear that many of these records are decidedly distance from the primary "positive" records and closer to being outliers. Of course, this may not remain the case in future efforts with an expanded data set.
4.3.2 Measuring Clustering Quality

By allowing for closer examination of the records, the case study addresses the fundamental issue of the visualizations efficacy being undercut by unreadable labeling along the x-axis. Without knowing which records are being clustered, there is no way of identifying whether the changes observed in the graphs indicate an increase or decrease in quality. For instance, consider the pre-course survey data removal, seen between Figures 6 and 7. According to the Source 3 features’ places in the rankings found in Section 3.1, they should be considered quite significant. However, given how these figures are presented, there is no way of identifying whether or not the records that were clustered in one figure but not another should be clustered. And yet, the changes between these figures are not so different from those seen between Figures 4 and 5. So, while the amount of information each feature carries can be estimated with these dendrograms, there is no verifiable evidence of clustering quality and certainly no metric to gauge it. Sections 3.3.2 and 3.3.3 address these two points, respectively, and, through the combination of their findings, provide empirical evidence.

4.3.3 Case Study: Elucidating Feature Shortcomings

With its tracking of the six "positive" records, the case study in Section 3.3.2 completely corroborates what the ML classifiers and feature rankings posit earlier. The records form a tight-knit group when closely following the feature rankings in their weights, but are scattered across the dendrogram in two different scenarios: when all features are present and equally weighted (as seen in Figures 12 and 13), and when the moderately significant Source 3 features are removed leaving only the GPA and Source 2 variables (see Figures 18 and 19).

The "negative" records present a more complex output in that there seems to be no discernible pattern in how the records’ positions change with the weight configurations.
Specifically, Figure 42 shows that, while modifying the weight configuration to match the Section 3.1 findings did bring a few of the records closer together, most of the "negative" cases remain scattered. However, their position in the initial uniform weighting (Figure 28) is worth noting. It is with this weight configuration that the records are most closely clustered when compared to their placements in any of the other dendrograms, although there is still substantial separation between them. What cohesion the records do have, however, completely breaks when the demographics and gender identity features are zeroed out. This can be interpreted to mean that out of all of the features collected for this study, these were the only ones that connected these records together. Again, the fact that these features are insignificant is indisputable based even on just the findings of this project alone, so it can only be concluded that if any comprehensive grouping of at-risk students is to be done, more features that specifically distinguish at-risk students need to be collected and reported. Recall the earlier conclusion from Section 4.1 about the need to explore the totality of external factors that affect students, rather than just a select few. This perfectly matches with the fact that "negative" cases are not being grouped together when only features such as survey responses and academic performance are being utilized. Both indicate a need for expanded context in order to better distinguish at-risk students so the models can more effectively identify them.

4.3.4 Confusion Matrices: An Urgent Need for More Data

Section 3.3.3’s confusion matrices also showcase this lacking ability to identify "negative" cases, but through the lens of the more practical, staff-usable scheme of labeling a cluster of at least a specific size as "positive" and the rest as "negative" (at-risk). It was when this cluster size parameter was 39 that the best results generally showed. Other thresholds were less effective; this could be because 39 is a median
value in a set that contains 79 records. A larger data set would be needed in order for this parameter’s effect to be known more conclusively. The very fact that the cluster size was (necessarily) parameterized leads to some issues with how the results can be compared with each other. Doing so is somewhat problematic because the cluster size often changes at the same time as the weight configuration. So, whether the changes in the true/false positive/negative counts and rates across figures are the result of modifying the feature vector or the figure’s particular cluster size cannot be absolutely confirmed. This ultimately invalidates the conclusions drawn from comparing figures that have more than one independent variable between them.

However, there are controlled comparisons available for closer inspection. Take, for example, Figures 46 and 48. Both have a cluster size of 39, but have different weight configurations; the former removes the demographics and gender identity while the latter also zeroes out the major and number of internships. The higher true negative count of Figure 48 is, of course, due to the change in feature selection causing the rearrangement of the records in the clustering pattern. So, it can be said that the cluster size alters the denominator of the true/false positive/negative ratios while the weight configuration controls the numerator. Further experimentation would be required to find a weight configuration that alters this "positive"/"negative" threshold more favorably, if one exists given this data set. It needs to be noted, however, that Figures 48, 56, and 63 all give the same (and best) results. Each has a cluster size of 39 but a different weight configuration, so it could be that a kind of level-off point for this data set has been reached for this particular scheme.

While on the topic of "negative" records, the limitations of this set and their effects need to be addressed. Table 12, which details all the confusion matrix results, includes not just the results from the figures in Section 3.3.3, which only use a grade threshold of 0.8, but also contains records of experiments done with grade thresholds of 0.78
and 0.87. Take note of the general increase in true negative rate in direct proportion
with the grade threshold. This is to be expected; if the grade threshold is higher,
there are more records labeled "negative". The true positive rate, however, takes a
major hit, going from an average percent in the mid-90s in the 0.78 and 0.8 grade
thresholds to ranging from 67 to 80 percent for the 0.87 threshold. It needs to be
understood, though, that these rates are not to be taken at face value. Consider the
0.87 threshold, which causes 23 out of the 79 records to be considered "negative",
many of which would not, in the real world, be considered as failing. Ultimately, what
Table 12 indirectly showcases is the most problematic portion of this entire study:
the fact that there is an extremely small number of records that describe "at-risk"
students. It is true that the features are the core reason why the models are not able
to effectively group these records. However, the lacking population is an obvious
contributing factor in that the few students that failed did so for different reasons.
In other words, these "negative" samples are not similar in terms of the features
available, meaning more features in which "at-risk" students are similar are definitely
required, but a deeper problem is that there is absolutely no guarantee that these
students will be clustered together. They could continue to be clustered instead with
"positive" students purely because the weighted Euclidean distance between those
records is the shortest available, in spite of whatever changes to the feature vector
are made. Recall that this issue is very similar to that of the ML classifiers. In that
scheme, the "negative" cases did not have a presence within the data set that was
substantial enough for the model to recognize them. Clearly, this logic does not only
apply to the clustering-based confusion matrices of Table 12, but to the supervised
learning-based binary and partition threshold schemes from Section 3.2 as well.

On a more final note, this study would be remiss to not acknowledge an ongoing
element of uncertainty. As concrete as these findings are, a greater knowledge of the
surrounding context (i.e. an expanded data set) can change, if only marginally, any of the results found in this study, including the ranking of features by significance or the records’ clustering patterns. Future efforts will need to acknowledge that the results and interpretations described above may change, although their being completely unfounded is unlikely.

4.4 A Reflection

As the author, I (Mustafa Emre Yesilyurt) would like to explain my reasoning behind taking on this study and what I think it means to myself and for others. This project was done in part to fulfill the graduation requirements for my degree, but at the forefront of my mind when accepting the offer was the prospect that my contributions would be of help to underserved students. I use the word "accepting" deliberately; there was no lengthy and agonizing decision process involved in going ahead with it. I took it on as soon as my advisor told me about it solely on the basis that I knew that, someday, it would help people. I could not say the same for the majority of the other subjects I had been studying as a possible graduate project seeing as they were largely centered around finding ways to improve existing technologies. I write this with no intention of besmirching those topics or those who would pursue them; I would even go so far as to argue that my being unable to find topics with as much of a human impact as what I ultimately chose was due to my own lack of imagination. Miraculously, that was circumvented by an incredible stroke of luck. This project all but fell onto my lap and I cannot be more thankful to those who worked to bring it to me. Regardless of how serendipitous the origins of my involvement with the project were, what I immediately understood was that it was a major responsibility. The responsibility was not for my own sake as every previous computer science assignment I attempted before had been. This was no longer just
about my qualifying to graduate, but of performing a service that would help increase
the chances of success for those who have the odds of achieving it stacked against
them.

UNVS 101 is a course with a genuine mission behind it: to help usher in an era of
economic and educational equality for the underserved, marginalized, and maligned.
This is the mission of which I had become a part that imbued me with a greater
value for the humanity at the heart of it. I think back to presenting my preliminary
findings on Student Research Day, remembering how I never left my poster because
I wanted to share with people the more human aspect that could not be expressed
only with the numbers on display. Now, I reflect even more frequently than I used to
on how many of my peers are alienated by the college experience, the difficulties it
presents for people of lower socioeconomic status that I never had to face and would
not wish on anyone. My efforts in this study to help augment the ability of UNVS 101
to reach more underserved people provides me some measure of consolation in each
of these pensive moments. But the conclusion to which I consistently return is that,
up to when I took on this project, I have been living as a part of an establishment
that, consciously or not, considers itself outright superior to these underserved peers
and can, has been, and will no doubt continue to grind them down by simply turning
away. There is a righteous anger at the heart of UNVS 101’s design and purpose;
it flies in the face of this oppressive establishment not just by refusing to neglect
students, but by offering support, guidance, and knowledge to those who want only to
be seen as equals and treated fairly i.e. with the barest minimum of respect. To have
contributed to a program that works to allow the underserved the same opportunities
as the privileged has been a true honor.

But I cannot ignore my status as an unwitting member of the establishment, and it is
upon understanding my own complicity in its oppression that I face a cruel truth. I
was given such a remarkable opportunity to be a part of this mission without lifting a finger to seek it out; this reads as a shameful act on my part. I cannot help but feel that, in completing my time on this project, I am leaving the mission behind to the detriment of an effort that only seeks to do right by people. This is a feeling that weighs on me, and I don’t know if it will become easier with time as I move toward concluding my education and officially starting my career. It is my sincere hope, however, that the findings in this study will prove helpful in aiding future efforts with UNVS 101. What I also hope for, perhaps even more so, is to reach other graduate students with the following message: their culminating experiences do not have to be in service of just the sciences. They can direct their efforts to help improve the lives of the people immediately around them. They can break away from the inherited pattern of neglecting our underserved peers. Their efforts can come from a place of humanity.

I have this hope of reaching those students because I believe that by standing in unity as a single student body, a single university campus, a single human species, we can work together to achieve feats that, more than being great, are beneficial, just, and kind.

I would like to thank my advisor, Dr. Teng Moh, and my committee members Dr. Melody Moh and Dr. Elaine Collins for their unwavering support during my time with this project. I look forward to where my career will take me and I hope that I will have more opportunities to build upon the contributions that I and many others have made toward achieving this mission. Thank you.
CHAPTER 5

Conclusion and Future Work

In indicating whether or not a student is on the path to success in UNVS 101, Figures 1, 2, and 3 clearly emphasize the students’ survey responses over whatever ethnic background they have or racial profile they fit. These findings are corroborated by the lackluster model performance in Tables 7 and 8 in tandem with the higher accuracy of the models in Tables 5, 6, 9, 10, and especially 11. Not only were the reported performances better for the latter tables, the affiliated standard deviations were often lower (demonstrating greater reliability), and the observed degradation in performance was usually much smaller than in the Source 1-only models. The PCA results seen in Tables 2, 3, and 4 dispute the corroborated feature rankings, but a closer inspection of the PCA technique revealed it to be unfit for the purposes of feature selection and not suited to the discrete data used in this study. It is also clear that any kind of model that is meant to predict a student’s performance needs to incorporate elements related to the students’ personal experiences and broader socioeconomic and academic status, not solely be based on their identities. In effect, more context, both in terms of features available and sample size, is required to generate higher-quality supervised-learning models.

This finding also holds true for the unsupervised learning model employed in this study. A single-linkage clustering algorithm using a weighted Euclidean distance function provided even further corroborating evidence of the feature rankings. Visualizations of clustering patterns seen in Section 3.3.1 showcased fewer cases of chain-link clustering when less significant features were pruned. The observed records in the scheme conducted in Section 3.3.2 demonstrated how records were grouped more effectively also as a result of those features’ removal. Additionally, Section 3.3.3’s confusion matrices found far more balanced true negative and true positive rates when
the feature weighting more closely matched Section 3.1’s rankings and only while utilizing a median cluster size threshold. Such balance in determining "negative" from "positive" records will prove instrumental for instructors and staff attempting to identify at-risk students by manually looking over their profiles. All of this is done by generating models that specifically remove the identity-based features that have been scientifically proven in the past to be insignificant and instead emphasizing the performance and survey-based ones that Section 3.1 highlighted.

Again, though, context is key in determining feature vectors. Those features related to identity can be useful if the overall feature vector more broadly encapsulates the individual’s place within a larger socioeconomic landscape, especially a student’s financial situation. While the collection of these kinds of data run the risk of violating individuals’ privacy, there is at least one that would not do this. A student’s status within SJSU’s Educational Opportunity Program (EOP), an initiative that is designed to provide further support for first-generation and otherwise disadvantaged students, was suggested as having the potential to be very useful to this project. Whether or not a student is involved with EOP could be a helpful indicator of where they are on the path to success, and it would be a welcome addition to the feature vector in future iterations of this project. Beyond a greater number of features, however, the data set also needs to be expanded vertically. There needs to be more than just 79 records with all of the survey responses completed and the final grade label available so that they can be used in training classifiers. Recall from the end of Section 4.3 how "negative" records were clustered with "positive" ones due primarily to a lack of other similar "negatives" with which they could be grouped. This idea of archetypal cases not being grouped with like cases due to a small population ties into the findings of the supervised-learning classifiers, where those same differing types of "negative" cases did not carry a presence large enough for the models to recognize them. This is
an issue that needs to be addressed for any substantial progress to be made in future efforts. The fact that this study’s findings are not as conclusive as they could be as a result of working with a smaller data set is a testament to this fact. In keeping with this uncertainty, it should be repeated that these results and feature rankings, as corroborated as they are, can be subject to change in the future pending expansion of the data set.

Regarding the techniques employed, the usage of classifiers and dependence on a truth label such as “grade” caused the performance degradation observed in Tables 5, 6, 7, 8, 9, 10, and 11. This happened when using thresholds closer to the average grade by forcing the separation of very similar records into being either a pass or fail, confusing the models. To address this problem, future studies may want to avoid classifiers (i.e. supervised learning techniques) and instead continue experimenting with clustering algorithms and other unsupervised methods. Again, though, this shift in technique would never address the core, fundamental issue of lacking data. That aside, many techniques (supervised and unsupervised) not tested in this study and may still warrant consideration.

With continued diligence in collection processes and cooperation from students, the data can grow into an even more encompassing and applicable training set to help UNVS 101 staff be able to reach even more at-risk individuals.
LIST OF REFERENCES


### LIST OF TABLES

1. Features in the data set used in this study. .................................. 10
2. PCA results for all available features. ................................. 22
3. PCA results for only Source 1 features. ................................. 23
4. PCA results for only Source 2 and 3 features. ............................. 23
5. Results of 10-fold cross-validation for binary threshold scheme using all features. ................................................. 25
6. Results of 10-fold cross-validation for partition threshold scheme using all features. .................................................. 25
7. Results of 10-fold cross-validation for binary threshold scheme using Source 1 features. .................................................. 25
8. Results of 10-fold cross-validation for partition threshold scheme using Source 1 features. .................................................. 26
9. Results of 10-fold cross-validation for binary threshold scheme using Source 2 and 3 features. ......................................... 26
10. Results of 10-fold cross-validation for partition threshold scheme using Source 2 and 3 features. ........................................ 26
11. Results of 10-fold cross-validation for partition threshold scheme using GPA, major, num_intern, and Source 2 and 3 features. 26
12. Confusion matrix results. .......................................................... 56
LIST OF FIGURES

1 Feature scores based on models trained with all available features (1-26). 19
2 Feature scores based on models trained with only non-survey features (1-14). 20
3 Feature scores based on models trained with only survey features (15-26). 21
4 Clustering with uniform weighting and all features present. 27
5 Uniformly weighted clustering without demographics and gender identity. 28
6 Uniformly weighted clustering without demographics, gender identity, major, or number of internships. Only GPA, weekly survey averages, number of weekly labs attended are present, and pre-course survey data. 29
7 Uniformly weighted clustering without demographics, gender identity, major, number of internships, or pre-course survey data. Only GPA, weekly survey averages, and number of weekly labs attended are present. 30
8 Clustering with uniform weighting without demographics, gender identity, major, number of internships, or number of weekly labs attended. Only GPA, weekly survey averages, and pre-course survey data Only GPA and weekly survey averages are present. 31
Uniformly weighted clustering without demographics, gender identity, major, number of internships, number of weekly labs attended, or pre-course survey data. Only GPA and weekly survey averages are present. 

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. GPA is weighted more to reflect its greater significance according to the feature rankings detailed in Section 3.1.

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1.

Clustering with uniform weighting and all features present. Positive cases are highlighted.

Clustering with uniform weighting and all features present. Positive cases are highlighted. Zoomed in.

Uniformly weighted clustering without demographics and gender identity. Positive cases are highlighted.

Uniformly weighted clustering without demographics and gender identity. Positive cases are highlighted. Zoomed in.

Uniformly weighted clustering without demographics, gender identity, major, or number of internships. Only GPA, weekly survey averages, number of weekly labs attended are present, and pre-course survey data. Positive cases are highlighted.
Uniformly weighted clustering without demographics, gender identity, major, or number of internships. Only GPA, weekly survey averages, number of weekly labs attended are present, and pre-course survey data. Positive cases are highlighted. Zoomed in.

Uniformly weighted clustering without demographics, gender identity, major, number of internships, or pre-course survey data. Only GPA, weekly survey averages, and number of weekly labs attended are present. Positive cases are highlighted.

Uniformly weighted clustering without demographics, gender identity, major, number of internships, or pre-course survey data. Only GPA, weekly survey averages, and number of weekly labs attended are present. Positive cases are highlighted. Zoomed in.

Clustering with uniform weighting without demographics, gender identity, major, number of internships, or number of weekly labs attended. Only GPA, weekly survey averages, and pre-course survey data. Only GPA and weekly survey averages are present. Positive cases are highlighted.

Clustering with uniform weighting without demographics, gender identity, major, number of internships, or number of weekly labs attended. Only GPA, weekly survey averages, and pre-course survey data. Only GPA and weekly survey averages are present. Positive cases are highlighted. Zoomed in.

Uniformly weighted clustering without demographics, gender identity, major, number of internships, number of weekly labs attended, or pre-course survey data. Only GPA and weekly survey averages are present. Positive cases are highlighted.
Uniformly weighted clustering without demographics, gender identity, major, number of internships, number of weekly labs attended, or pre-course survey data. Only GPA and weekly survey averages are present. Positive cases are highlighted. Zoomed in.

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. GPA is weighted more to reflect its greater significance according to the feature rankings detailed in Section 3.1. Positive cases are highlighted.

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. Positive cases are highlighted. Zoomed in.

Clustering with uniform weighting and all features present. Negative cases are highlighted.
29 Clustering with uniform weighting and all features present. Negative cases are highlighted. Zoomed in.

30 Uniformly weighted clustering without demographics and gender identity. Negative cases are highlighted.

31 Uniformly weighted clustering without demographics and gender identity. Negative cases are highlighted. Zoomed in.

32 Uniformly weighted clustering without demographics, gender identity, major, or number of internships. Only GPA, weekly survey averages, number of weekly labs attended are present, and pre-course survey data. Negative cases are highlighted.

33 Uniformly weighted clustering without demographics, gender identity, major, or number of internships. Only GPA, weekly survey averages, number of weekly labs attended are present, and pre-course survey data. Negative cases are highlighted. Zoomed in.

34 Uniformly weighted clustering without demographics, gender identity, major, number of internships, or pre-course survey data. Only GPA, weekly survey averages, and number of weekly labs attended are present. Negative cases are highlighted.

35 Uniformly weighted clustering without demographics, gender identity, major, number of internships, or pre-course survey data. Only GPA, weekly survey averages, and number of weekly labs attended are present. Negative cases are highlighted. Zoomed in.
Clustering with uniform weighting without demographics, gender identity, major, number of internships, or number of weekly labs attended. Only GPA, weekly survey averages, and pre-course survey data Only GPA and weekly survey averages are present. are present. Negative cases are highlighted. 

Clustering with uniform weighting without demographics, gender identity, major, number of internships, or number of weekly labs attended. Only GPA, weekly survey averages, and pre-course survey data Only GPA and weekly survey averages are present. are present. Negative cases are highlighted. Zoomed in.

Uniformly weighted clustering without demographics, gender identity, major, number of internships, number of weekly labs attended, or pre-course survey data. Only GPA and weekly survey averages are present. Negative cases are highlighted.

Uniformly weighted clustering without demographics, gender identity, major, number of internships, number of weekly labs attended, or pre-course survey data. Only GPA and weekly survey averages are present. Negative cases are highlighted. Zoomed in.

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. GPA is weighted more to reflect its greater significance according to the feature rankings detailed in Section 3.1. Negative cases are highlighted.
Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. GPA is weighted more to reflect its greater significance according to the feature rankings detailed in Section 3.1. Negative cases are highlighted. Zoomed in.

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. Negative cases are highlighted.

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. Negative cases are highlighted. Zoomed in.

Confusion matrix of clustering with uniform weighting and all features present. "Positive" cluster size is 40.


Uniformly weighted clustering without demographics and gender identity. "Positive" cluster size is 39.

Uniformly weighted clustering without demographics, gender identity, major, or number of internships. Only GPA, weekly survey averages, number of weekly labs attended are present, and pre-course survey data. "Positive" cluster size is 28.
48 Uniformly weighted clustering without demographics, gender identity, major, or number of internships. Only GPA, weekly survey averages, number of weekly labs attended are present, and pre-course survey data. "Positive" cluster size is 39.

49 Uniformly weighted clustering without demographics, gender identity, major, number of internships, or pre-course survey data. Only GPA, weekly survey averages, and number of weekly labs attended are present. "Positive" cluster size is 27.

50 Uniformly weighted clustering without demographics, gender identity, major, number of internships, or pre-course survey data. Only GPA, weekly survey averages, and number of weekly labs attended are present. "Positive" cluster size is 46.

51 Clustering with uniform weighting without demographics, gender identity, major, number of internships, or number of weekly labs attended. Only GPA, weekly survey averages, and pre-course survey data Only GPA and weekly survey averages are present. "Positive" cluster size is 26.

52 Clustering with uniform weighting without demographics, gender identity, major, number of internships, or number of weekly labs attended. Only GPA, weekly survey averages, and pre-course survey data Only GPA and weekly survey averages are present. "Positive" cluster size is 45.

53 Uniformly weighted clustering without demographics, gender identity, major, number of internships, number of weekly labs attended, or pre-course survey data. Only GPA and weekly survey averages are present. "Positive" cluster size is 34.
Uniformly weighted clustering without demographics, gender identity, major, number of internships, number of weekly labs attended, or pre-course survey data. Only GPA and weekly survey averages are present. "Positive" cluster size is 50.

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. GPA is weighted more to reflect its greater significance according to the feature rankings detailed in Section 3.1. "Positive" cluster size is 26.

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. GPA is weighted more to reflect its greater significance according to the feature rankings detailed in Section 3.1. "Positive" cluster size is 39.

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 13.

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 15.

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 22.
Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 26. 74

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 32. 75

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 33. 76

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 39. 77

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 46. 78

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 50. 79
Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 56.

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 61.

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 65.

Clustering with only GPA, weekly survey averages, number of weekly labs attended, and pre-course survey data present. Both GPA and weekly survey averages are weighted more to reflect feature rankings detailed in Section 3.1. "Positive" cluster size is 70.