



Classifying Major Choice Through Grades and Academic Interest

Napat Sakulsaengprapha

**Capstone Final Report for BSc (Honours) in
Mathematical, Computational and Statistical Sciences**

Supervised by: Dr. Timothy Wertz

AY 2021/2022

Yale-NUS College Capstone Project

DECLARATION & CONSENT

1. I declare that the product of this Project, the Thesis, is the end result of my own work and that due acknowledgement has been given in the bibliography and references to ALL sources be they printed, electronic, or personal, in accordance with the academic regulations of Yale-NUS College.
2. I acknowledge that the Thesis is subject to the policies relating to Yale-NUS College Intellectual Property (Yale-NUS HR 039).

ACCESS LEVEL

3. I agree, in consultation with my supervisor(s), that the Thesis be given the access level specified below: [check one only]

☒ Unrestricted access

Make the Thesis immediately available for worldwide access.

☐ Access restricted to Yale-NUS College for a limited period

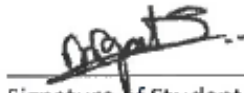
Make the Thesis immediately available for Yale-NUS College access only from _____ (mm/yyyy) to _____ (mm/yyyy), up to a maximum of 2 years for the following reason(s): (please specify; attach a separate sheet if necessary):

After this period, the Thesis will be made available for worldwide access.

☐ Other restrictions: (please specify if any part of your thesis should be restricted)

Napat Sakulsaengprapha, Elm College

Name & Residential College of Student



Signature of Student

30/3/2022

Date



Name & Signature of Supervisor

30/3/2022

Date

Acknowledgements

While my college experience was not what I expected, due to COVID-19, I would not have it any other way. The past 4 years have been challenging but also a time of growth for me. I would like to first thank all the professors who have been kind and patient in educating me. As Yale-NUS College, as we know, is going to be gone after 2025, I hope this capstone can act, to our community, as a small reminder of what Yale-NUS was like.

This capstone would not have been possible without Professor Tim Wertz who has been kind, prompt, and encouraging throughout the process. Having covered 30MC under Prof Wertz's tutelage (from QR to Capstone), I am proud to say that I am graduating with a minor in Tim Wertz. Jokes aside, thank you very much.

To all my friends (you know who are), thank you for the fun memories. Special shoutout to my suitemates (Cavan, Mark, Nich, Tavis, Zhifeng) for making Elm feel like home.

Lastly, but most importantly, this opportunity to study in Singapore would not have been possible in the first place without my family. Thank you for your endless support and love.

Abstract

B.Sc (Hons)

Classifying Major Choice Through Grades and Academic Interest

by Napat SAKULSAENGPRAPHA

Upon applying to Yale-NUS College, students are asked to indicate their academic interests. After completing the Common Curriculum in their first and second year, students then declare their major. While some people's academic interests correspond directly to what they end up majoring in, others change their mind and pick a different major. In this capstone, various machine learning models are trained using academic interest data and Common Curriculum grade data to classify students into the 14 majors offered. Using both academic interest data and Common Curriculum grade data, students were successfully classified into their majors at an accuracy of 22.74% (AdaBoost), 32.49% (Random Forest), 34.66% (Multinomial Logistic Regression), 35.01% (Decision Tree), 36.10% (Gradient Boosting), and 36.82% (Support Vector Machine). Overall, the highest accuracy was 37.18% which was achieved when using Support Vector Machine trained on only the academic interest data.

Keywords: Data Science, Machine Learning, Classification, SVM, AdaBoost, Random Forest, Decision Tree, Multinomial Logistic Regression, GBM

Word Count: 9080

Table of Contents

Chapter 1: The Introduction	6
1.1 Background	6
1.2 Motivating the Problem	7
Chapter 2: The Data	8
2.1 Academic Interests and Major	8
2.2 Grade Scales	8
2.3 Missing Shadow Grades	9
2.4 Common Curriculum Review	10
Chapter 3: Imputing Missing Data	12
3.1 Multiple Multivariate Regression	12
3.2 Multivariate Imputation by Chained Equations	14
Chapter 4: Exploratory Data Analysis	17
4.1 Grade, Academic Interest, and Major Distributions	17
4.2 Circular Transition Visualization with Circlize Package	19
Chapter 5: Models	21
5.1 Benchmark	21
5.2 Training with Only Grade and Academic Interest Data	22
5.3 Individual Major Binary Logistic Regression Classification	23
5.4 Multinomial Logistic Regression Classification	24
5.5 Decision Trees and Random Forests	25
5.6 Gradient Boosting Classification and AdaBoost	27
5.7 Support Vector Machine	28
5.8 Classification by Division	29
5.9 Natural Groups: K-means, PCA, Hierarchical Clustering	30
Chapter 6: Discussions	36
6.1 Results	36
6.2 Limitations	37
Bibliography	38
Appendix A:	41
A.1 Code	41
A.2 Literature Review	41
A.3 Supplementary and Additional Figures	44
A.4 R Packages Used For Certain Functions and Algorithms	51

Chapter 1: The Introduction

1.1 Background

At Yale-NUS College, there are many stories about how someone who initially started college with a particular interest ends up majoring in an unrelated field. Upon applying to Yale-NUS College, applicants are asked to pick two academic areas of interest. Students are to declare their major only at the end of their second year. One standard experience all students have at Yale-NUS is the Common Curriculum, a set of multi-interdisciplinary courses that “will introduce [students to] multiple modes of inquiry, and to some of the timeless ideas of human existence.”¹ The goal of this capstone project is to develop machine learning models that will be able to classify students’ majors at Yale-NUS by using Yale-NUS Class of 2018-2023’s academic interest and Common Curriculum module performance data. The results consist of models which predict a multiclass variable (major) from a mix of quantitative (e.g., module grades) and categorical (e.g., academic interest) data. Data cleaning, exploratory data analysis, and modeling are all performed using R.

¹<https://www.yale-nus.edu.sg/academics/overview/academic-experience/>

1.2 Motivating the Problem

The original motivation behind this capstone is based on applying the findings to improve the college experience for future Yale-NUS cohorts. Because this capstone project is based on a case study of Yale-NUS, the capstone's direct application is limited due to the announced merger of Yale-NUS and the NUS University Scholars Programme. However, findings and models developed in this capstone can be applied to other educational institutions that have a similar Common Curriculum or major system.

One concrete benefit of how this study could be applied is to find a more efficient way of assigning academic advisors. Finding whether or not initial academic interest has a significant effect on major choice can lead to more compatible student-academic advisor pairing; if significant, the College may adopt the idea of assigning faculty from the student's declared academic interest as the advisor.

While universities hold large data on students, little systematic work has been done on predicting a student's actual major choice. However, the few studies on this topic have suggested similar findings that academic interest is a better predictor of major than grades. One study by Stahmann at the University of Utah found that intended division of registration was a better predictor of major than academic achievement (Stahmann, 1969). Similarly, a more contemporary study by Allen and Robbins found that while the difference may be small, interest-major composite is a better predictor of major persistence than first-year GPA (Allen and Robbins, 2007). More details on literature reviews can be found in the [Appendix](#).

Chapter 2: The Data

2.1 Academic Interests and Major

When applying to Yale-NUS College, applicants are asked to declare two areas of academic interests. Unlike other universities, Yale-NUS College students do not have a major or faculty they are attached to once they matriculate. Students are able to freely choose from 14 majors when declaring their academic interests (when applying) and major (at the end of their second year). One exception is the Double-Degree Programme in Law and Liberal Arts (DDP). DDP is excluded from the analysis because a student cannot declare DDP as their academic interest and cannot simply choose to major in DDP because admission into DDP is contingent on a separate application to NUS Law as well. Thus, this capstone will only focus on the transition from academic interest to major of the 14 majors at Yale-NUS.

2.2 Grade Scales

Yale-NUS College uses a 5.0 point grading scale (Yale-NUS College, 2021).

Grades	A+	A	A-	B+	B	B-	C+	C	D+	D	F
Numeric Points	5.0	5.0	4.5	4.0	3.5	3.0	2.5	2.0	1.5	1.0	0.0
Adapted Points	5.5	5.0	4.5	4.0	3.5	3.0	2.5	2.0	1.5	1.0	0.0

Figure 2.1: Letter Grades and Grade Points at Yale-NUS

The data available are students' Common Curriculum grades which are all in letter grades. For the purpose of subsequent analyses, the grades are quantified by changing them to numeric points. For this capstone, Yale-NUS's letter grades conversion to numeric points grades are adapted slightly, changing the numeric points associated with an "A+" to a 5.5. The 0.5 increment was chosen to be in line with the 0.5 difference between other grade steps. This is done in order to differentiate students who performed exceptionally well in the Common Curriculum modules; there were only 59 instances of A+ in 11,160 grades across the Common Curriculum grades of the 6 cohorts. In the case that a certain module significantly influences a decision for a certain major, the differentiation between A and A+ can be useful in classification.

The Common Curriculum modules of interest are: Comparative Social Inquiry, Literature and Humanities 1, Literature and Humanities 2, Modern Social Thought, Philosophy and Political Thought 1, Philosophy and Political Thought 2, Quantitative Reasoning, and science courses². Historical Immersion is excluded as most students take this course after declaring their major.

2.3 Missing Shadow Grades

During the first semester of the first year, students will receive comments and grades so that they can familiarize themselves with collegiate work. Instead of letter grades, students will receive either a Completion Satisfactory (CS) or Completion Unsatisfactory (CU). Both a shadow grade (letter grade) and the published grade (CS/CU) was recorded from the Class of 2020 onwards. Shadow grades are letter grades that are not recorded on transcripts but represent students' performances. However, the

² There have been many versions of the science courses, so we have not specified the names here.

Class of 2018 and Class of 2019 only had the published grade recorded. Thus, the dataset is missing grade data for the Class of 2018 and 2019 in their first semester of year 1. Before training classification models, attempts to impute these missing grades for the Class of 2018 and 2019 is first made. The grade data for the Class of 2020 and above are systematically complete.

2.4 Common Curriculum Review

As Yale-NUS is a relatively young institution, the Common Curriculum constantly undergoes review; changes were made to both the content and timing of the courses. As a result of these changes, science courses grades taught in year 1 semester 1 for the Class of 2018 and 2019 are missing. Science courses grades are available for the Class of 2020 and beyond.

More notably, the content of science courses have had significant changes. There have been 9 versions of the science courses³. Many of the reviews have led to drastically different content within the courses; hence, the different iterations cannot be viewed as equivalent. Because the data for the Class of 2018 and 2019's first science course is missing and the experience across the different cohorts was very different for the science courses, imputation of missing shadow grades for the Class of 2018 and 2019's science course is not attempted. Instead, this capstone congregates the available grades for the science courses into one numeric value. This value will represent the students' overall performance in science. For example, a student in the Class of 2018 has taken three science courses⁴, where two grades are available and one is missing, will be assigned one

³ Scientific Inquiry, Scientific Inquiry 1, Scientific Inquiry 2, Foundations of Science, Foundations of Science 1, Foundations of Science 2, Integrated Science 1, Integrated Science 2, Integrated Science 3

⁴ 2 of which were half-semester courses

numeric grade, which is the average of the two available grades. For later cohorts where shadow grades are available, they are assigned an average grade based on the science courses they have taken.

With the simplification of the science grades, which is subsequently referred to in congregate as Science, the missing grade data that will be imputed for the Class of 2018 and 2019 are for Literature and Humanities 1, Philosophy and Political Thought 1, and Comparative Social Inquiry. One advantage of imputing the data over dropping them is having a large sample to train and test the models on. Dropping the data for the Class of 2018 and 2019 would mean losing data points from around 300 students. Further, because LH1, PPT1, and CSI are related to LH2, PPT2, and MST respectively, there is a sound basis in believing that imputation can be reasonably accurate. The pre-imputed data can be seen in [Figure A.3](#).

Chapter 3: Imputing Missing Data

This chapter deals with imputing missing shadow grade data for the Class of 2018 and 2019 in their first semester of college. The aim is to predict the grades for Literature and Humanities 1, Philosophy and Political Thought 1, and Comparative Social Inquiry using the rest of the Common Curriculum modules grade data.

3.1 Multiple Multivariate Regression

“Multivariate Multiple Regression (MMR) is the method of modeling multiple responses, or dependent variables, with a single set of predictor variables” (Ford, 2017). Breaking down the name of this method, “MMR is multiple because there is more than one IV [Independent Variable] [and] is multivariate because there is more than one DV [Dependent Variable]” (Datallo, 2013). MMR regresses each of the response variables separately on the predictors, as if separate regressions are ran for each dependent variable. The difference between a normal regression is that a modified hypothesis tests for parameters and confidence intervals for predictions must be used (Ford, 2017).

In order to perform MMR on the data, data is separated into two sets: those in the Class of 2018 and 2019 (older) and those in the later batches (newer). A model is trained using the newer data; subsequently, this model is used to predict the grades for the older cohorts. In order to determine whether or not to include all the predictors in a MMR, a multivariate test statistics must be used. Among the multiple MANOVA tests such as the Wilk’s Lambda, Lawley - Hotelling Trace, Roy’s Largest Roots, there is evidence that the Pillai’s Trace is the most robust test “with adequate power to detect true differences

in a variety of situations” (Olson, 1974). Further, the Pillai Trace is most robust when there are departures from MANOVA assumptions, which are: (Bray and Maxwell, 1985)

1. Units are randomly sampled from the population of interest,
2. Observations are statistically independent,
3. The dependent variables have a multivariate normal distribution within each group,
4. The k groups have a common within group population covariance matrix.

Using the Pillai test statistic as the determinant of whether or not to include a predictor, the results suggested that all 5 other Common Curriculum modules should be included in the model.

```

Type II MANOVA Tests: Pillai test statistic
      Df test stat approx F num Df den Df    Pr(>F)
LH2      1  0.052538   10.7944      3    584 6.529e-07 ***
PPT2      1  0.059325   12.2770      3    584 8.471e-08 ***
MST       1  0.021397    4.2564      3    584 0.005478 **
QR        1  0.071439   14.9768      3    584 2.102e-09 ***
Science   1  0.064696   13.4653      3    584 1.659e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 3.1: MANOVA (Pillai Trace) Tests Result for Imputation Model

Applying our model to the older batch by inputting their 5 other Common Curriculum grades allows us to predict their performance in Literature and Humanities 1, Philosophy and Political Thought 1, and Comparative Social Inquiry. To see if the model’s result is representative, a plot of the predicted grade distribution and actual grade distribution (from other years) is created. As can be seen in [Figure A.5](#) in the appendix, the variation of the distribution of the predicted grades is much lower than that of the actual grades. This is not surprising as MMR is a relatively simple imputation method that assumes linearity. Due to the substantial variance reduction, another method is subsequently explored.

3.2 Multivariate Imputation by Chained Equations

Multivariate Imputation by Chained Equations (MICE) is a commonly used package in R that uses the distribution of the observed data/variables to approximate multiple possible values for the missing data points (Katitas, 2019). Further, MICE is a useful method because it “allows [us] to account for the uncertainty around the true value, and obtain approximately unbiased estimates [which] allows us to calculate standard errors around estimations” (Katitas, 2019). In steps, MICE works as follows (Kropko et al., 2014):

1. Select values that keep the relationship in the dataset intact in place of missing values,
2. Generate independently drawn imputed (default = 5) datasets,
3. Calculate new standard errors using variation across datasets in order to reflect uncertainty from the imputed dataset.

One assumption that MICE takes is that the data is missing at random (MAR) - which means that the propensity of missingness depends on the observed data, not the missing data (Rubin, 1976). While this dataset’s missingness stems from a systematic fact of missing shadow grades, it still makes sense to assume that missing values can be replaced by predictions derived from the observable portion of the dataset (other grades) because the missing modules all have a related counterpart. While there are various methods that can be used in MICE for imputing data, the method that is most suitable for this analysis is PMM. It is suitable because PMM is used for numeric variables and that it “is the only method that yields plausible imputations and preserves the original data distributions ... [and is] independent from the missingness mechanism” (Vink et al.,

2014). This resolves both issues of preservation of original data distribution that occurred when using the multiple multivariate regression and the possible violation of the MAR assumptions.

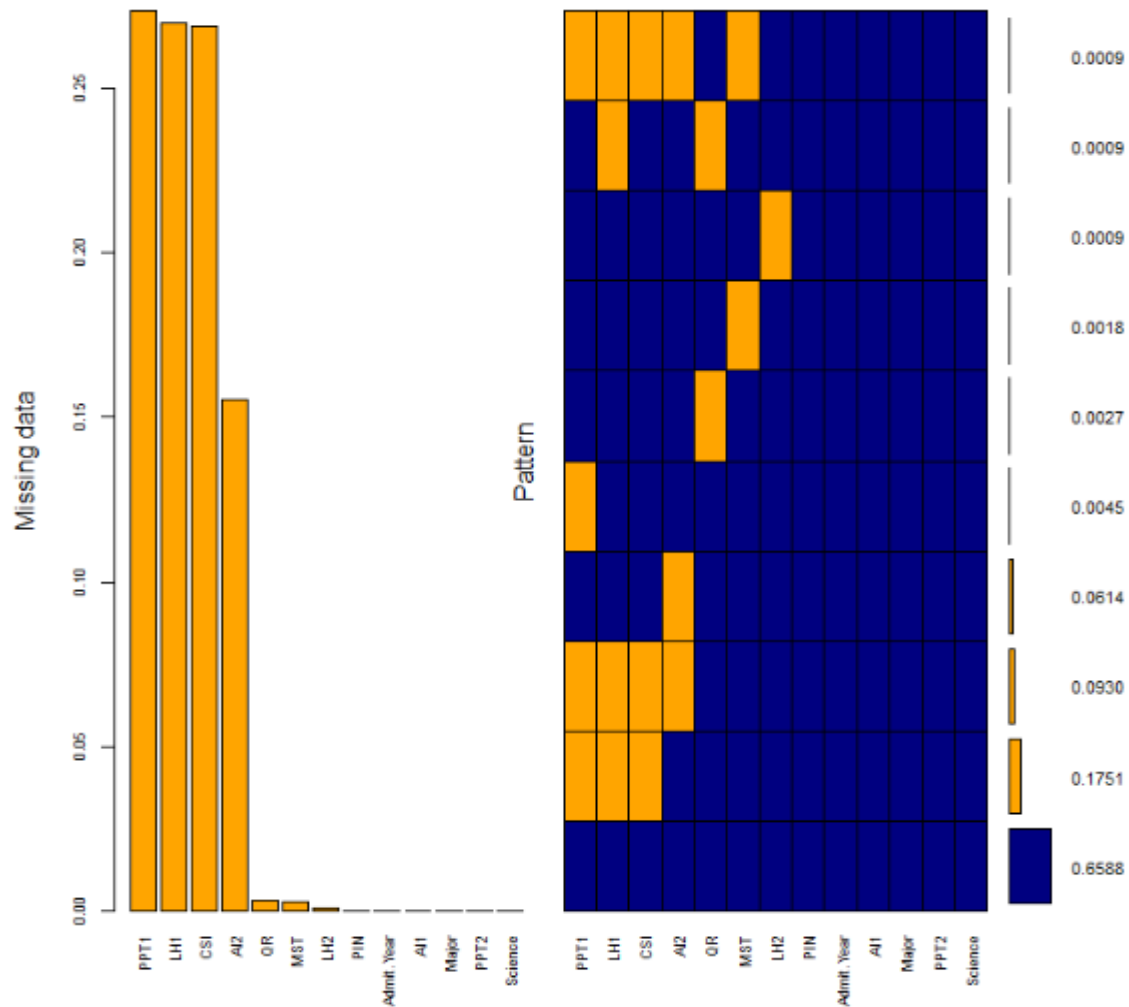


Figure 3.2: Missing Data Pattern

Before imputing the data, it is useful to visualize and understand the pattern of the missingness of the data. This can be achieved by using the function *aggr()* from the package VIM in R. A quick interpretation of Figure 3.2 would be that 65.58% of the values in the data set has no missing values, 17.51% of the values have missing data for PPT1, LH1, and CSI, and 9.30% of the values have missing data for PPT1, LH1, CSI,

and AI2. The majority of the missing data stems from not having the shadow grades for the first semester for the Class of 2018 and 2019. AI2 data is missing as some students did not indicate a secondary area of academic interest upon applying.

After implementing the imputation using MICE, the comparison of the distribution of the predicted grades and the actual grades in Figure 3.3 shows that they are quite similar. Because MICE improved upon MMR, MICE predictions are used to complete the dataset. The post-imputed dataset can be seen in Figure [A.4](#).

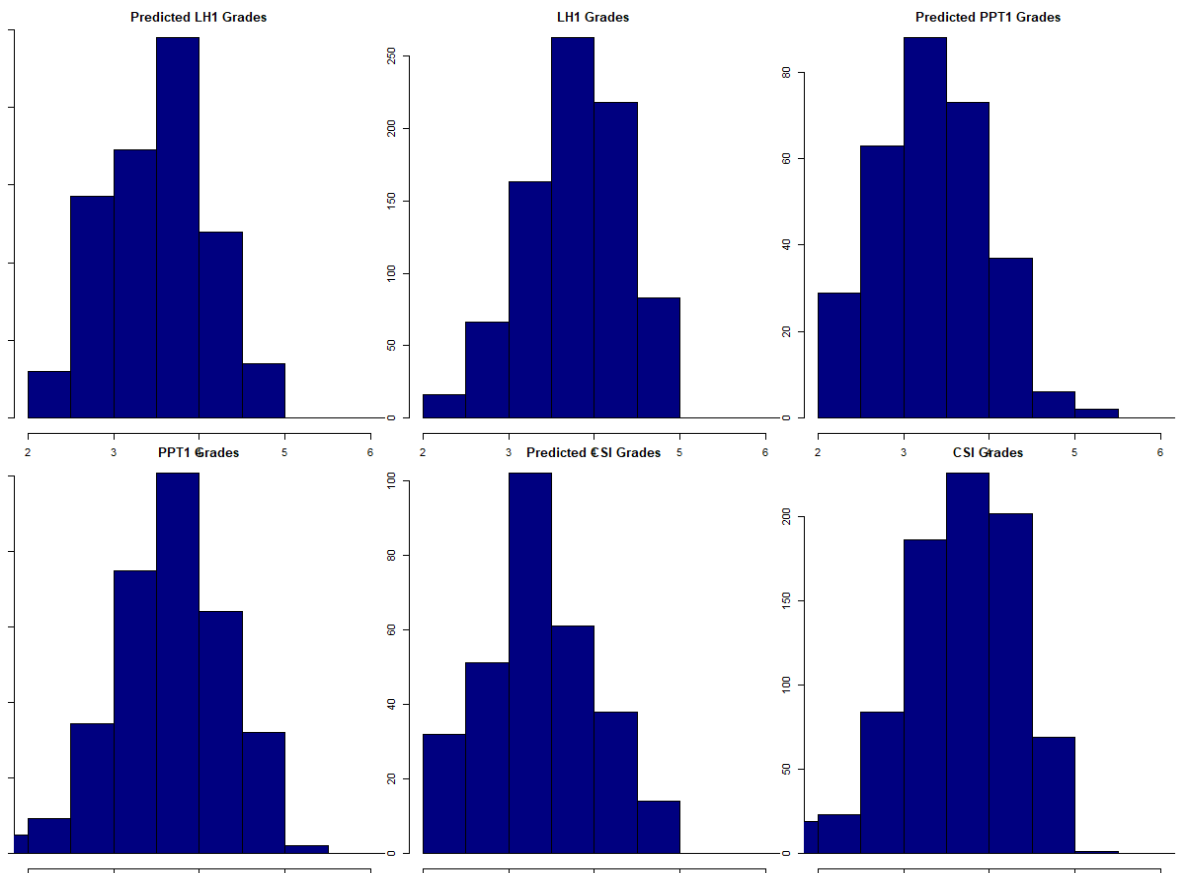


Figure 3.3: Distribution of Imputed Data vs. Actual Data (MICE PPM)

Chapter 4: Exploratory Data Analysis

This chapter features various graphs and diagrams that are beneficial for summarizing the data's main characteristics.

4.1 Grade, Academic Interest, and Major Distributions

As can be seen in [Figure A.6](#) in the appendix, most Common Curriculum modules' grades are centered around 4, which corresponds to a B+. Visually, QR and Science, which are skewed left, have higher variance than other Common Curriculum modules. The variance of QR is 0.61 and the variance of Science is 0.90; these values are much larger than MST's variance which is 0.27.

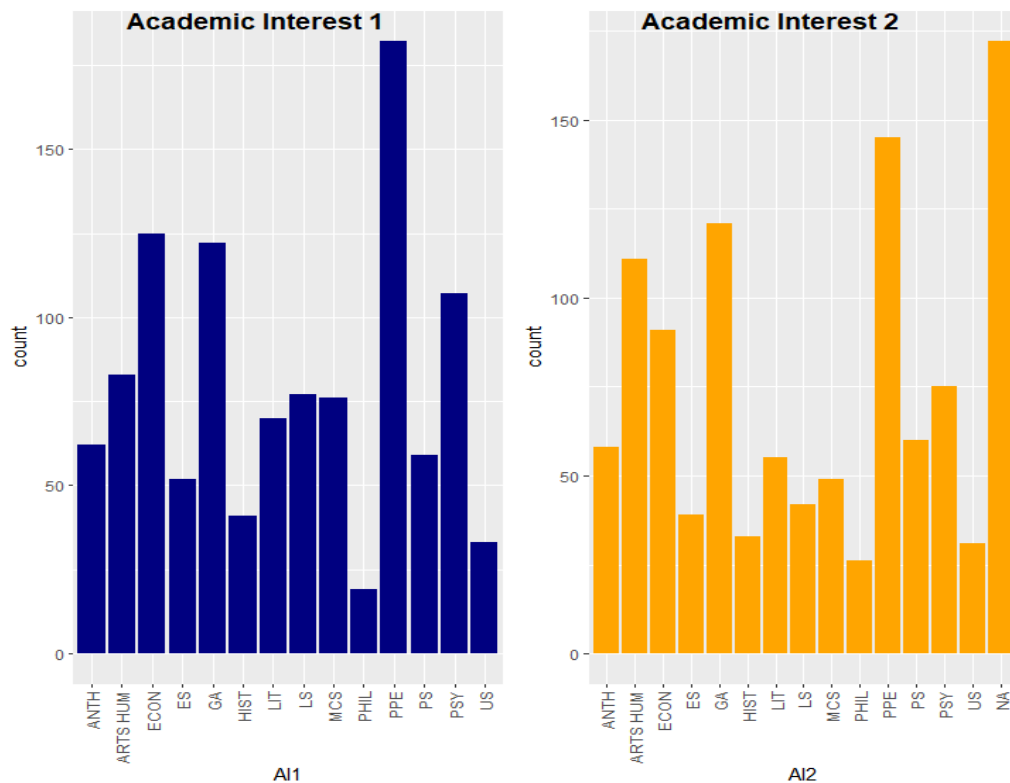


Figure 4.1: Academic Interest Popularity

The most popular Academic Interests [1] are PPE, ECONS, and GA respectively. NA was picked as the most common value for Academic Interest 2. Similar to Academic Interest [1] the most popular area of academic interest is PPE; GA and ARTS HUM came second and third respectively.

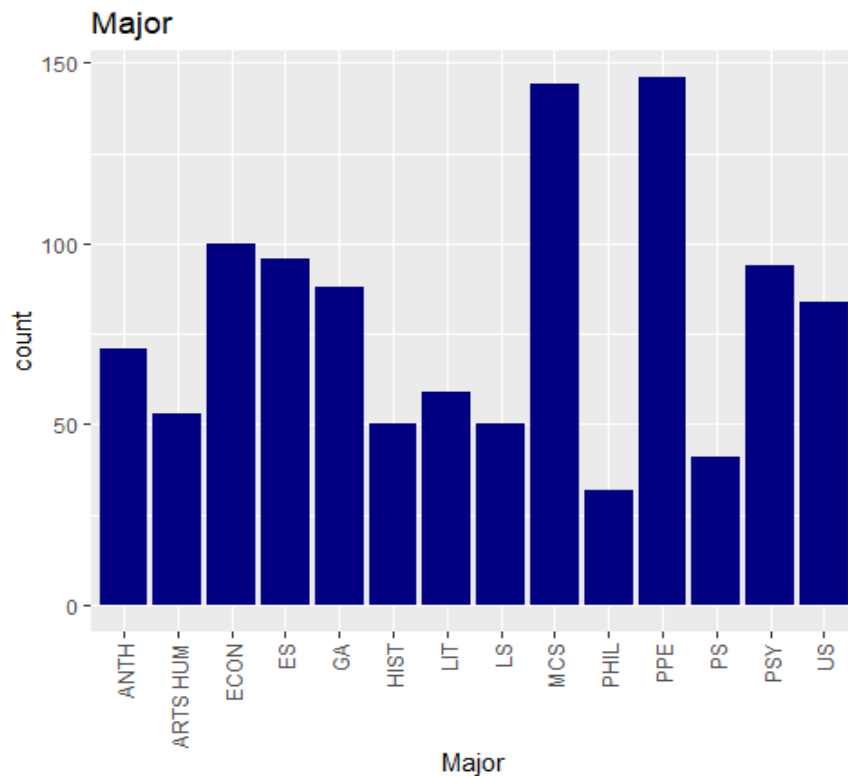


Figure 4.2: Major Distribution

The top two most popular majors that students declared are MCS and PPE, as can be seen in Figure 4.2. While PPE is a major that garnered a lot of interest since application, MCS was ranked seventh in terms of the number of people who stated that it was their academic interest. This reaffirms that there was indeed movement between academic interests upon applying and majors declared at the end of year 2 for students at Yale-NUS. This movement is the motivation behind the models that will be subsequently trained to classify students' majors.

4.2 Circular Transition Visualization with Circlize Package

Visualizing the transitions between academic interest to major can offer valuable insight to understanding some common underlying trends. However, traditional methods like a network, with nodes and edges, is not useful as the 14 majors would result in extremely messy graphs which reveal little information. Alternatively, one could use the *chordDiagram()* function in the circlize package to create a circular transition visualization. Gu, who is the author of the circlize package, claims that “circular layout is very useful to represent complicated information [as]... it elegantly represents information with long axes or a large amount of categories ... shows data with multiple tracks focusing on the same object ... [and] provides an efficient way to arrange information on the circle” (Gu, 2014).

Prior to using the visualization, an adjacency matrix must be created from the data. An adjacency matrix is a matrix in which the “value in the i^{th} row and the j^{th} column represents the relation from object in the i^{th} row and the object in the j^{th} column where the absolute value measures the strength of the relation” (Gu, 2014). In [Figure A.7](#), where ES represents the 2nd row and PPE represents the 1st column, the value of 2 in $m_{2,1}$ means that 2 people who initially declared ES as their Academic Interest [1], declared PPE as their major. The adjacency matrix can be seen in [Figure A.7](#) in the appendix. Using this adjacency matrix, the chordDiagram in Figure 4.3 is formed.

Academic Interest to Major Transitions

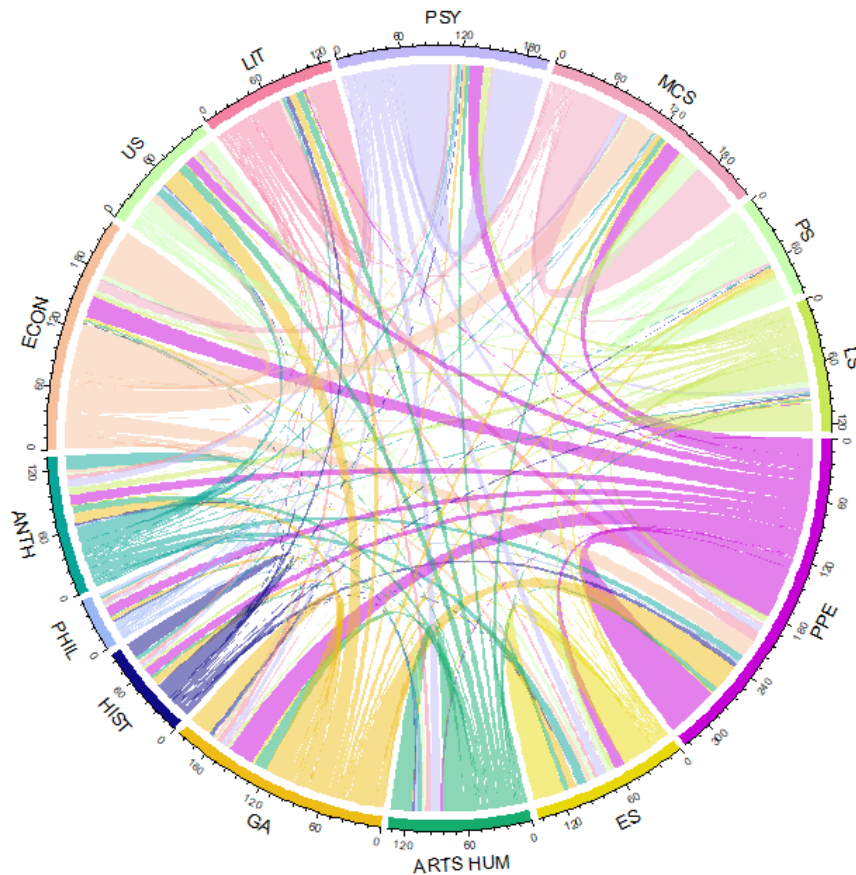


Figure 4.3: *chordDiagram* for All to Major Transition

This circular visualization reveals multiple stories. For example, there are more people who initially declared ECON as their academic interest that end up in MCS than the opposite way around: the beige colored line moving from ECON to MCS is thicker than the light-pinked colored line moving from MCS to ECON. Further, the amount of people who initially declared PPE and majored in PPE is less than those who initially declared other majors as their academic interests and ended up majoring in PPE; this can be seen by the fact that the neon-pinked line moving from PPE back to itself is less than the sum of every other colored line moving to PPE. These are some examples of the insights that the *chordDiagram* can provide. Next, models to classify majors are created.

Chapter 5: Models

This chapter focuses on fitting several supervised and unsupervised machine learning models to classify students into majors. In supervised learning, models are trained to classify students into the 14 discrete majors by giving it labels (major) along with other data points such as Common Curriculum grades and academic interest. On the other hand, unsupervised data takes unlabeled data (Common Curriculum grades and academic interest) to identify natural groups. This chapter describes the difficulties faced and how alternative methods and solutions are considered to solve the problem at hand. The dataset of 1108 students is split into a training set (75% of the data) and a test set (25% of the data), which is used to verify the predictive accuracy of the models.

5.1 Benchmark

To answer the question of what a good accuracy is, it is necessary to first define a benchmark. One possible benchmark to use is random chance: since there are 14 majors, predicting the correct major of a student based on random chance alone would be $1/14$ or 7.14%. However, a more appropriate benchmark is to use Academic Interest 1's direct conversion to Major as a benchmark. This is simply assigning students' majors based purely on Academic Interest 1. This is a more challenging benchmark to beat as it tests how much additional predictive power Academic Interest 2 and the Common Curriculum grades have in determining student's majors or how efficiently Academic Interest 1 can be used by the models to classify majors. This second benchmark which was calculated from the test set to be 34.65% will be the main benchmark that is subsequently referenced.

5.2 Training with Only Grade and Academic Interest Data

First, Common Curriculum performance and academic interest data are used to train the models separately to classify students' majors. This isolates the effects these two factors have on predictive accuracy. A more detailed explanation of these models will be discussed subsequently when both Common Curriculum grades and academic interest are used jointly in training the models.

Method	Accuracy Common Curriculum Grades Only	Accuracy Academic Interests Only
AdaBoost	8.66%	22.11%
Random Forest	14.44%	35.74%
Multinomial Logistic Regression	17.32%	35.74%
Gradient Boosting	17.33%	33.94%
Decision Tree	17.69%	35.01%
SVM (Sigmoid)	17.69%	37.18%

Figure 5.1: Accuracy of Models Trained Solely with Grade vs. Academic Interest Data⁵

In general, it can be clearly seen in Figure 5.1 that training the models with only the academic interest data resulted in a much higher accuracy than using just grade data. The accuracy achieved with Support Vector Machine (SVM) when trained only with academic interest is the highest at 37.18%. This result matches the results in the literature reviews by Stahmann and Allen & Robbins that academic interest is a better predictor of major than grades. In the following sections, both academic interest and grade data are used jointly to classify students' majors, attempting to improve accuracy.

⁵ Green represents beating the benchmark while red represents worse performance than the benchmark. Dark green and dark red indicate the best and worst accuracy respectively. This applies to all similar tables.

5.3 Individual Major Binary Logistic Regression Classification

Before attempting to classify students into the 14 available majors using the entire dataset, it is reasonable to build a basic foundation, which is to classify binarily whether or not a student would belong to each major. The logistic regression predicts the likelihood of an outcome based on input variables; the logistic regression is based on the logistic function $f(y)$ and y , which is expressed as a linear function of the input variables $(x_1 \dots x_{p-1})$ (Dietrich et al., 2015):

$$f(y) = \frac{e^y}{1+e^y} \text{ for } -\infty < y < \infty$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{p-1} x_{p-1}$$

The logistic regression utilizes the *logit* link function, which is also known as the log odds ratio (Dietrich et al., 2015). Using p to denote $f(y)$:

$$\ln\left(\frac{p}{1-p}\right) = y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{p-1} x_{p-1}$$

When solving for p , “0.5 is [commonly] used as the default probability threshold to distinguish between any two class labels” (Dietrich et al, 2015). Maximum Likelihood Estimation (MLE) is a technique used to estimate the model parameters. In this binary classification of major example, a student with p , probability, greater than the threshold would be classified as being in the particular major and not if lower. To optimize the prediction of whether a student belongs to a particular major, the threshold used was also varied to find the threshold that produced the highest accuracy in the test set for each major.

	major	Accuracy	Threshold
1	MCS	0.877256317689531	0.55
2	LS	0.96028880866426	0.4
3	PSY	0.916967509025271	0.5
4	PPE	0.888086642599278	0.6
5	ES	0.927797833935018	0.25
6	PS	0.974729241877256	0.75
7	GA	0.906137184115523	0.45
8	HIST	0.967509025270758	0.75
9	ANTH	0.949458483754513	0.45
10	ECON	0.902527075812274	0.45
11	LIT	0.963898916967509	0.75
12	ARTS HUM	0.953068592057762	0.85
13	US	0.916967509025271	0.85
14	PHIL	0.981949458483754	0.55

Figure 5.2: Accuracy of Binary Logistic Regression and Optimal Threshold

Figure 5.2 reveals that the accuracy to predict each major is relatively high, ranging from 87.72% for MCS to 98.19% for Philosophy. The optimal threshold also had quite a high variation ranging from 0.25 to 0.85, with the majority between 0.40 and 0.60. One interpretation for the lower accuracy for MCS is that there are a lot of movements into and out of the major that are influenced by other factors beyond grades and initial interest such as job prospects and demand for computer science and AI.

5.4 Multinomial Logistic Regression Classification

Building on the results of the binary logistic regression classification, a logical step forward is to classify students into the 14 majors directly rather than whether they are in a certain major or not. The method used is the multinomial logistic regression classification which is simply an extension of the binary logistic regression that allows for more than two classes of outcome variable. It similarly uses MLE to evaluate probability of belonging to certain classes. Multinomial Logistic Regression is a

favorable method as it “does not assume normality, linearity, or homoscedasticity” (Starkweather and Moske, 2011). Both Common Curriculum grades and academic interest data were used to predict the major.

	ANTH	ARTS	HUM	ECON	ES	GA	HIST	LIT	LS	MCS	PHIL	PPE	PS	PSY	US
334	0.00		0.00	0.14	0.00	0.01	0.00	0.02	0.00	0.70	0.02	0.04	0.05	0.00	0.03
1084	0.08		0.00	0.08	0.08	0.13	0.05	0.00	0.00	0.09	0.03	0.27	0.00	0.11	0.07
387	0.00		0.00	0.11	0.00	0.01	0.00	0.06	0.00	0.67	0.02	0.06	0.04	0.00	0.03
980	0.01		0.01	0.22	0.06	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.04	0.49
332	0.01		0.00	0.06	0.05	0.01	0.00	0.00	0.05	0.52	0.02	0.00	0.25	0.02	0.00
1064	0.13		0.07	0.01	0.12	0.08	0.01	0.00	0.00	0.01	0.00	0.03	0.00	0.52	0.03

Figure 5.3: Example of Multinomial Logistic Regression Probability Outputs

Each row in Figure 5.3 displays the probability that the model assigns the particular student to be in each major. For example, Student 334 has 0.14, 0.01, 0.02, 0.70, 0.02, 0.04, 0.05, 0.03 probability to major in Economics, Global Affairs, Literature, MCS, Philosophy, PPE, and Urban Studies respectively. The model would then predict this student to be in MCS as it has the highest probability of 0.70.

[Figure A.8](#) is the confusion matrix produced when the trained multinomial logistic regression model is used to predict the major in the testing set. The numbers on the diagonal represent correct predictions made by the model. Summing the diagonal and dividing by the total number of data in the test set yields an accuracy of 34.66%, which only beat the benchmark by 0.01%.

5.5 Decision Trees and Random Forests

Decision trees are supervised machine learning algorithms which adopt a set of rules to make classification decisions. The intuition for decision trees is using data features to train a model by constructing test points and branches: each test point checks a certain input variable and each branch represents the decision being made. (Dietrich et

al., 2015). Traversing down branches represents the decision being made - the final point in the tree represents the prediction of the tree (Dietrich et al., 2015). The algorithm ideally continues this splitting process until the leaf nodes are pure, meaning that all data belong to a single class. However, it is common that the trees built end up having mixed leaf nodes - multiple classes in a single leaf node. In this case, the algorithm would assign the most common class of the particular node to the observation of the test set as the prediction (Bento, 2021). An illustration can be seen in [Figure A.10](#).

The ideal tree is a tree with the fewest splits possible that can accurately classify all data points; at each split, the algorithm minimizes the loss function that is based on the purity of the resulting nodes such as the Classification Error Rate, Gini Index or Cross-Entropy (Le, 2018). The model yielded the following confusion matrix upon testing it with the test set:

	predict_major														
	ANTH	ARTS	HUM	ECON	ES	GA	HIST	LIT	LS	MCS	PHIL	PPE	PS	PSY	US
ANTH	0		2	0	1	0	0	2	1	0	0	6	0	1	1
ARTS HUM	0		6	1	0	0	0	0	0	0	0	3	0	2	3
ECON	0		1	13	1	0	0	0	1	2	0	9	0	0	1
ES	0		1	3	7	0	0	0	1	2	0	5	0	3	2
GA	0		3	1	0	0	0	1	0	1	0	10	0	1	2
HIST	0		0	0	0	0	0	2	0	0	0	9	0	1	1
LIT	0		0	0	0	0	0	6	1	0	0	6	0	0	0
LS	0		1	0	0	0	0	0	6	2	0	1	0	4	0
MCS	0		3	10	0	0	0	1	1	17	0	5	0	4	1
PHIL	0		1	2	0	0	0	1	0	0	0	2	0	1	0
PPE	0		3	4	0	0	0	2	0	1	0	21	0	2	1
PS	0		0	0	1	0	0	0	0	4	0	2	0	0	0
PSY	0		3	0	0	0	0	2	3	2	0	3	0	17	0
US	0		4	3	0	0	0	0	1	0	0	5	0	0	4

Figure 5.4: Confusion Matrix of Decision Tree Prediction vs. Actual Major

As can be seen by the consecutive vertical zeros in certain columns in Figure 5.4, the decision tree did not predict any students would end up in Anthropology, Global Affairs, History, Philosophy and Physical Sciences. This reveals a short-coming of the decision tree: if it splits the data in even more nodes, it runs the risk of overclassification. The accuracy achieved by the decision tree is 35.01%, a slight

improvement upon the multinomial logistic regression classification, but still barely beating the benchmark of 34.65%.

Another commonly used methodology in classification is the random forest classifier. Random forests produce numerous trees and combine their results to form predictions. It mainly uses two techniques. Firstly, random forests utilize bagging, where “an individual tree is built on a random sample of the dataset ... this is repeated dozens or hundreds of times and the results are averaged” (Lesmeister, 2015). While the trees grown are not pruned based on error measure, their high variance is overcome by averaging the results (Lesmeister, 2015). Secondly, random forest takes a random sample of the input features at each split (Lesmeister, 2015). An illustration can be seen in [Figure A.11](#). One improvement upon the decision tree is that the model actually classified at least one student into each major, unlike the decision tree. This reflects the natural advantage multiple trees have over one single decision tree: certain majors that are not predicted in a particular tree are classified by another tree. The accuracy yield by the random forest classifier was 32.49%. This is rather a surprising result as this is both lower than the benchmark and accuracy of decision trees.

5.6 Gradient Boosting Classification and AdaBoost

Because the trees in random forests are built independently, they are unable to learn from the mistakes of other trees. A logical step forward is to attempt using algorithms that could learn from their mistakes to form better predictors. Both Gradient Boosting Classification and Adaptive Boosting (AdaBoost) are supervised machine learning models which use iterative ensemble methods. The intuition behind boosting methods, in general, is that an initial model is built, residuals are examined, and a new

model is fit based on these residuals around the loss function; this process is repeated until a certain stop criterion is met (Lesmeister, 2015). Gradient Boosting aims to increase the performance of the model by fitting new trees on the previous trees' residuals; newer trees are added to the ensemble which begins with a single weak learner, usually a decision tree with only a few splits (Boehmke and Greenwell, 2020). AdaBoost, on the other hand, trains subsequent models by reweighting the weak learners that underfit to improve accuracy (Murphy, 2012).

The accuracy of predictions achieved by AdaBoost and Gradient Boosting are 22.74% and 36.10% respectively. It is rather surprising that AdaBoost underperformed Gradient Boosting by almost 14% in terms of accuracy. AdaBoost employs a specific loss function, exponential loss, while Gradient Boosting is a generic algorithm that aids in searching for approximate solutions; this makes Gradient Boosting more flexible in application than AdaBoost (Choudhury, 2021). Thus, it is possible that the exponential loss function is not suitable in this application of classifying majors.

5.7 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm that is commonly used in classification problems. The intuition of SVM is simple: the algorithm finds the optimal line or hyperplane that subdivides the data in such a way that the margin, the distance between the separators, is maximized, increasing the probability that a new observation will fall on the correct side and classified correctly (Lesmeister, 2015). An illustration can be seen in [Figure A.12](#) in the appendix.

When the data at hand is of a high dimension and consequently not linearly separable, transformation is needed. Kernel transformation enables modeling

nonlinearity by efficiently expanding the feature space with the goal of achieving an approximate linear separation (Boehmke and Greenwell, 2020). Because Common Curriculum grades and academic interest data are high dimensional, kernel transformation is needed. The kernels that were considered are sigmoid, polynomial, and radial basis. The model trained that achieved the highest accuracy upon testing used the sigmoid kernel and achieved an accuracy of 36.82%, which is the highest out of all algorithms used. However, this accuracy still only beat the benchmark by 2.16% and underperformed training SVM with only academic interest data, which is accurate to 37.18%. For SVM, adding grade data to the training process reduced the accuracy of the model.

5.8 Classification by Division

Having seen that the various machine learning algorithms are unable to beat the accuracy benchmark for classification by much, it makes sense to explore whether prediction of groups of majors would yield a better result. In this case, an existing group of majors is the divisions in which Yale-NUS classifies majors into. The natural intra-division movement from Academic Interest [1] to Major is 59.56%, which is the benchmark for classification by division. Upon applying the algorithms, Figure 5.5 reveals the accuracy results:

Method	Accuracy Common Curriculum Grades Only
Multinomial Logistic Regression	51.62%
AdaBoost	54.24%
Gradient Boosting	59.21%
SVM (Sigmoid)	59.57%
Random Forest	60.29%
Decision Tree	62.45%

Figure 5.5: Accuracy of Models Trained to Classify Major Divisions

Half of the machine learning models underperformed the benchmark while the other half outperformed. Decision trees gave the highest accuracy, predicting 62.45% of the students' division correctly. However, decision trees only outperformed the benchmark by 2.89%. A question that arises is whether there are other natural groupings of majors that will allow these machine learning models to classify better.

5.9 Natural Groups: K-means, PCA, Hierarchical Clustering

The most basic way to start finding natural groups and clustering them is to use K-means clustering. K-means clustering is an unsupervised algorithm which, for a chosen value of k , a number usually based on contextual knowledge, “identifies k clusters based on the objects' proximity of the center of the k groups” (Dietrich et al., 2015). While K-means is a valuable method for finding natural clusters, its weakness comes from having to specify the number of k clusters. In addition to the 14 possible k from the separate major and the 3 k from the divisions of major, the number of k can be defined using the elbow method when no assumption or knowledge about the

classification problem is available. The elbow method requires plotting the Within Sum of Squares (WSS) against the number of clusters in order to find the “elbow” of the WSS curve, which is the number of clusters when the reduction of WSS becomes fairly linear (Dietrich et al., 2015). In the analysis shown in Figure 5.6, K-means was performed using k being equal to 3 (divisions), 4 (elbow method) and 14 (majors).

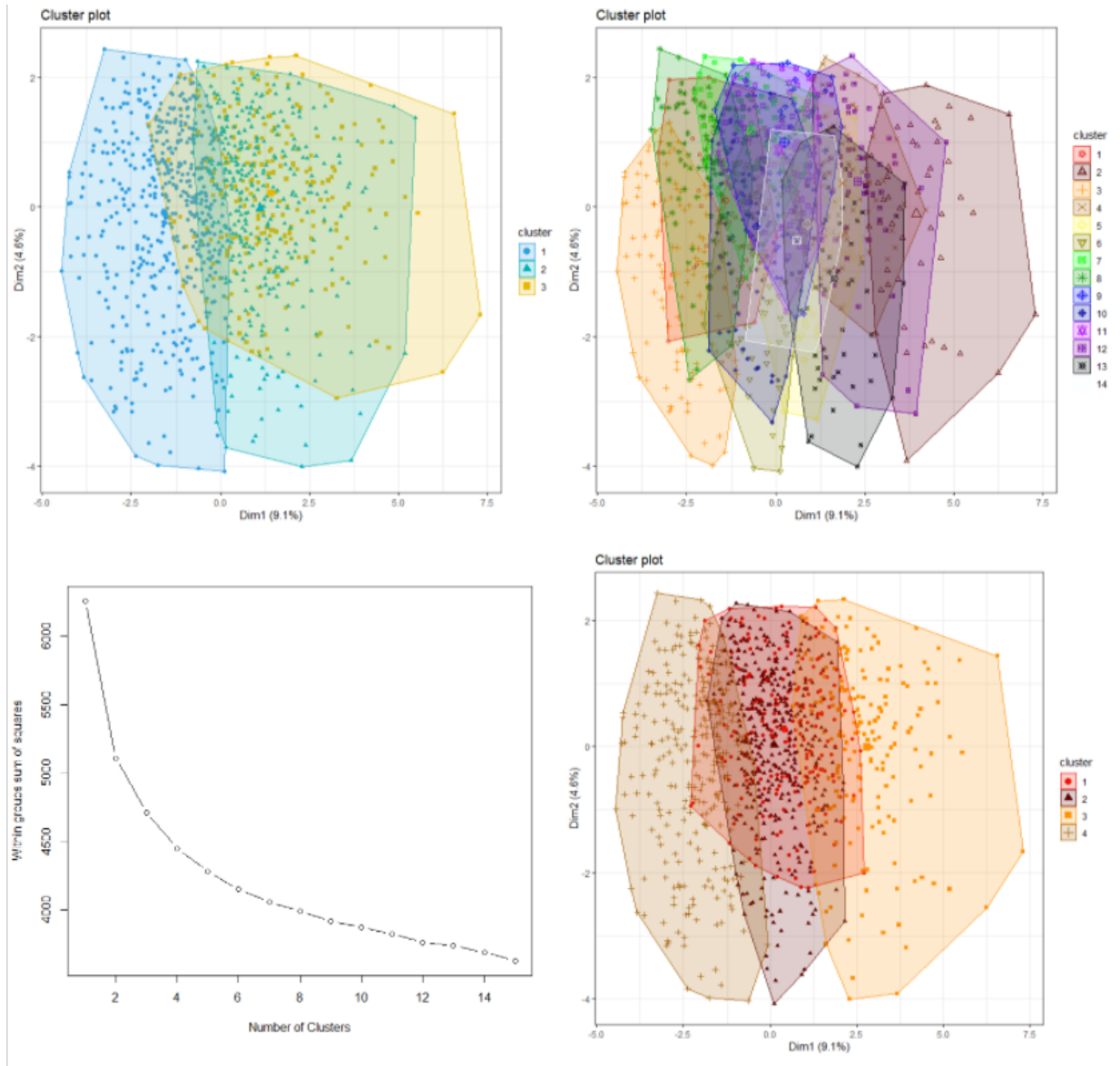


Figure 5.6: K-Mean Graphs

As can be seen in all 3 K-means plots in Figure 5.6, the formed clusters have large overlaps, indicating that using $k = 3, 4, 14$ does not allow for efficient clustering in 2-dimensions. An alternative method that is simple to implement for finding natural clusters is Principal Component Analysis (PCA). PCA is an unsupervised machine learning technique that attempts “to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set ... achieved by transforming [old variables] to a new set of variables, the principal components (PCs), which are uncorrelated, and ... the first few retain most of the variation” (Jolliffe, 2002). The first and second principal components can be plotted in 2 dimensions, where clusters, which aren’t visible due to high dimension, could be seen if they exist. Unfortunately, the 2 dimensional plots of the first and second principal component displayed in [Figure A.13](#) do not reveal any significant clusters. Upon adding labels to the plot, there are still too many overlaps for any useful information to be extracted regarding natural groups of majors.

Seeing that both K-means and PCA were not able to yield new potential natural clusters of majors in 2 dimensions, a final attempt is made in trying to find natural groups. A more complex clustering method that requires minimal human decisions is preferred. Upon searching for algorithms that can cluster into an unknown number of clusters, hierarchical clustering was chosen. Agglomerative hierarchical clustering will be the method of choice as packages in R allow clustering with minimal human intervention. Agglomerative hierarchical clustering starts each observation as a single cluster and successively merges clusters together until a criterion is reached (Pathak, 2018). The algorithm works in the following way (Pathak, 2018):

1. Calculate distance between every pair of observation to add to a distance matrix,

2. Puts all points in its own cluster,
3. Merge closest pairs of points based on distances from the distance matrix, reducing number of clusters by 1,
4. Recomputes distance between new clusters, store new information in new distance matrix,
5. Repeat steps 2 and 3 until there is only 1 cluster.

The output of a hierarchical cluster is a dendrogram. Usually, the optimal number of clusters is found by finding the longest unbroken line in the dendrogram, creating a vertical line at that point, and counting the number of crossed lines (Baeldung, 2020). Because the dataset contains over 1100 observations, it is hard to identify the longest unbroken branch. Luckily, the function *cutreeDynamic()* from the package *dynamicTreeCut* can assist in automatically finding the optimal number of clusters through an iterative process of cluster decomposition and combination which stops when the number of clusters becomes stable (Langfelder et al., 2007). The result can be seen in Figure 5.7.

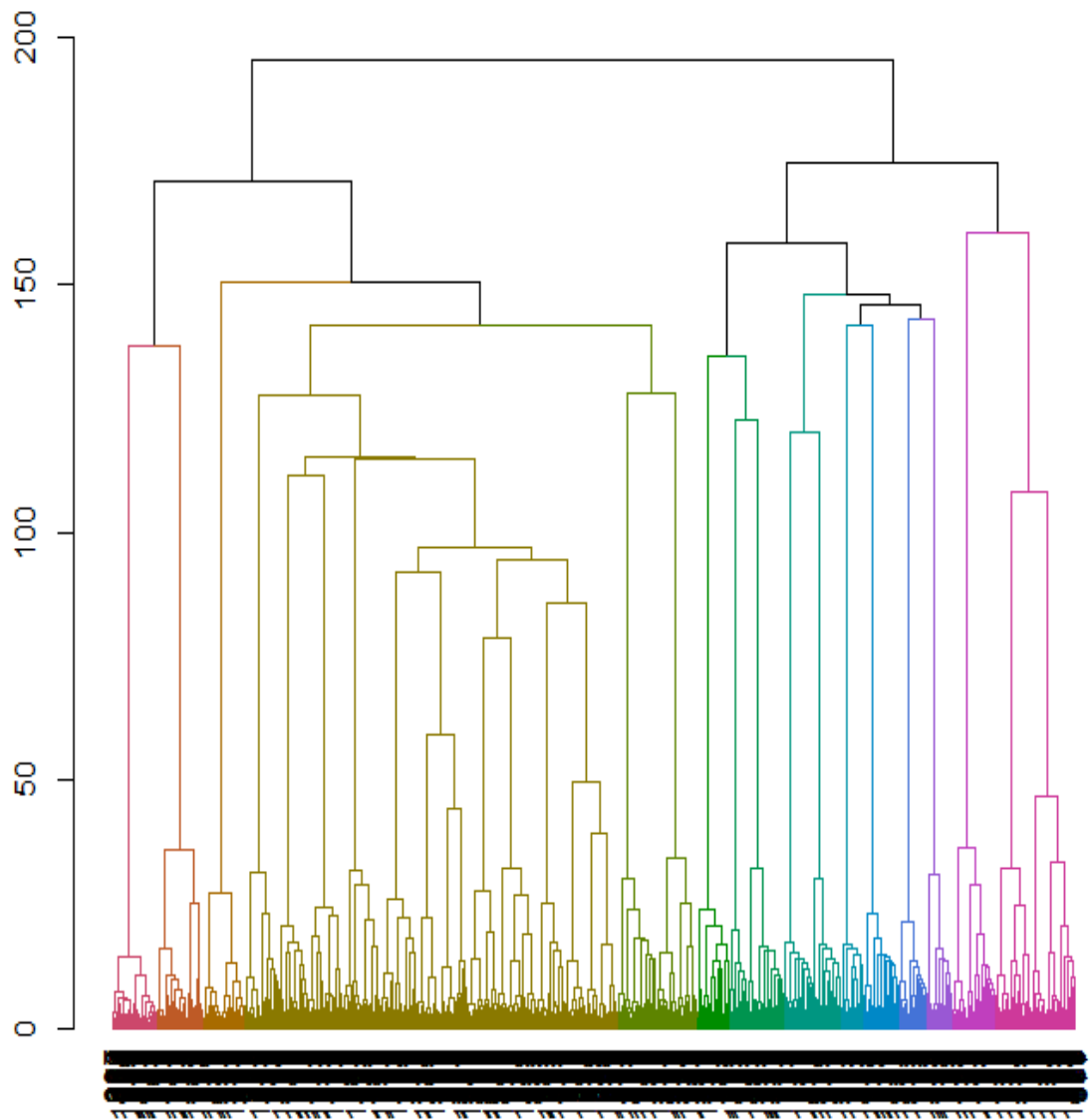


Figure 5.7: Agglomerative Hierarchical Clustering Dendrogram Clusters

The lowest number of clusters formed, upon trying all the possible distance calculation methods, when using agglomerative hierarchical clustering is 14. This result used the “ward.D2” method, Wards Method, that aims to minimize the variance within clusters (University of Alberta, n.d.). This is not an ideal result as it does not meaningfully reduce the number of classes from the 14 majors.

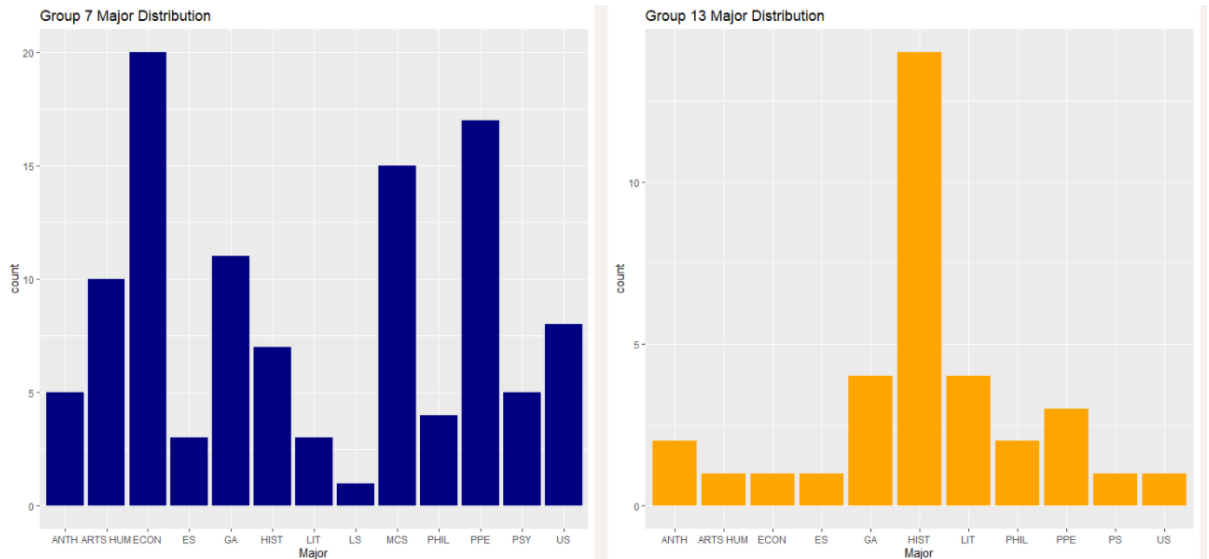


Figure 5.8: Sample of Clusters Found by Agglomerative Hierarchical Clustering

Figure 5.8 reveals two clusters, out of the 14, found by agglomerative hierarchical clustering. The graph on the left indicates a group which has a large population of Economics, MCS, and PPE majors. While there is a strong signal of Economics, MCS, and PPE majors, these 3 large majors also appear strongly in many other groups, which suggests that the appearance of these students in these groups may be due to the fact that these are particularly large majors. On the other hand, the graph on the right shows a strong membership of History majors, which is a relatively small major at Yale-NUS. This suggests that History majors have academic interest and Common Curriculum grades which is relatively distinguishable from other majors. Overall, most of the clusters formed did not have a clear division of majors. Because Hierarchical Clustering did not reduce the number of natural groups of majors below 14, further classification of natural groups were not explored.

Chapter 6: Discussions

6.1 Results

Using both academic interest data and Common Curriculum grade data, students were successfully classified into their majors at an accuracy of:

Method	Accuracy Common Curriculum Grades Only
SVM	36.82%
Gradient Boosting	36.10%
Decision Tree	36.10%
Multinomial Logistic Regression	35.01%
Random Forest	32.49%
AdaBoost	22.74%

Figure 6.1: Accuracy of Models Trained to Classify Major Divisions Using All Data

Upon training the models with academic interest data and Common Curriculum data separately, Common Curriculum grades achieved an accuracy between 8.66% to 17.69% while academic interest yielded accuracy rates ranging from 22.11% to 37.18%. Overall, SVM was the best classifier in this capstone project. All models beat random chance (7.14%) of guessing students' major. However, the better-performing models only beat the direct conversion of Academic Interest 1 to Major by only a few percentage points. Ultimately, training the machine learning models with academic interest and Common Curriculum grade data only slightly improved the accuracy rate of predicting students' majors from simply guessing their major by converting the declared academic interest.

6.2 Limitations

There are a few limitations and reasons that could have impaired the success of this Machine Learning project's classification results:

- The data set is imbalanced. “Imbalanced data typically refers to a problem with classification problems where the classes are not represented equally” (Brownlee, 2015). In Figure 4.3, it is evident that certain majors have a very high number of students while others have much fewer. One way to combat imbalance data is to resample - oversample under-represented class or undersample over-represented class (Brownlee, 2020). Upon trying both oversampling and undersampling - the overall accuracy dropped for all models except AdaBoost. However, the increase in accuracy of AdaBoost was not large enough to change the overall conclusions.
- The Common Curriculum experience is not captured fully by purely the grade data. While grades may be one factor that pushes students towards certain majors, there is a myriad of other information in the Common Curriculum itself that can potentially have a huge impact on selecting majors. Some examples include which professor they had, the professors' area of expertise, classmate interaction, and lack of content coverage for certain majors.
- The data also omitted other variables, beyond the Common Curriculum, which could significantly affect students' choice of major. Some examples include: what elective did students choose before declaring their major, whether or not they did well in those electives, what academic interests do their suitemates and friends have, and which major does the job market demand at their time. Adding these variables could be one possible extension of this study or future similar studies.

Bibliography

- Allen, Jeff, and Steven B. Robbins. "Prediction of College Major Persistence Based on Vocational Interests, Academic Preparation, and First-Year Academic Performance." Springer Netherlands, August 4, 2007.
<https://link.springer.com/article/10.1007/s11162-007-9064-5#citeas>.
- Baeldung. "Clustering Into an Unknown Number of Clusters." Baeldung on Computer Science, October 20, 2020.
<https://www.baeldung.com/cs/clustering-unknown-number>.
- Bento, Carolina. "Decision Tree Classifier Explained in Real-Life: Picking a Vacation Destination." Towards Data Science, July 18, 2021.
<https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575>.
- Boehmke, Bradley, and Brandon Greenwell. *Hands-on Machine Learning with R*. Boca Raton (Fla.): CRC Press, 2020.
- Bray, James H., and Scott E. Maxwell. *Multivariate Analysis of Variance*. Newbury Park: Sage, 1985.
- Brownlee, Jason. "8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset." Machine Learning Mastery, August 19, 2015.
<https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/>.
- Choudhury, Ambika. "AdaBoost vs Gradient Boosting: A Comparison of Leading Boosting Algorithms." Analytics India Magazine, January 18, 2021.
<https://analyticsindiamag.com/adaboost-vs-gradient-boosting-a-comparison-of-leading-boosting-algorithms/>.
- Dattalo, Patrick. "Multivariate Multiple Regression." Oxford Scholarship Online. Oxford University Press, 2013.
<https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199773596.001.0001/acprof-9780199773596-chapter-4>.
- Dietrich, David, E. Heller, and Beibei Yang. *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Indianapolis, IN: Wiley, 2015.
- Dunn, Frances E. "Two Methods for Predicting the Selection of a College Major." American Psychological Association. American Psychological Association, 1959. <https://psycnet.apa.org/record/1960-06553-001>.

- Ford, Clay. “Getting Started with Multivariate Multiple Regression.” University of Virginia Library, October 27, 2017.
<https://data.library.virginia.edu/getting-started-with-multivariate-multiple-regression/>.
- Gu, Zuguang. “Circular Visualization in R.” A Bioinformatician, 2014.
https://jokergoo.github.io/circlize_book/book/index.html.
- Jolliffe, I. T. *Principal Component Analysis*. New York: Springer, 2002.
- Katitas, Aycan. “Getting Started with Multiple Imputation in R.” University of Virginia Library, May 1, 2019.
<https://data.library.virginia.edu/getting-started-with-multiple-imputation-in-r/>.
- Kropko, Jonathan, Ben Goodrich, Andrew Gelman, and Jennifer Hill. “Multiple Imputation for Continuous and Categorical Data: Comparing Joint Multivariate Normal and Conditional Approaches.” *Political Analysis* 22, no. 4 (2014): 497–519. <http://www.jstor.org/stable/24573085>.
- Langfelder, Peter, Bin Zhang, and Steve Horvath. “Defining Clusters from a Hierarchical Cluster Tree: the Dynamic Tree Cut Package for R.” *Bioinformatics* 24, no. 5 (November 16, 2007): 719–20.
- Le, James. “R Decision Trees Tutorial: Examples & Code in R for Regression & Classification.” DataCamp, June 19, 2018.
<https://www.datacamp.com/community/tutorials/decision-trees-R>.
- Lesmeister, Cory. *Mastering Machine Learning with R Master Machine Learning Techniques with R to Deliver Insights for Complex Projects*. Birmingham: Packt, 2015.
- “Mice: Mice: Multivariate Imputation by Chained Equations.” RDocumentation. Accessed October 31, 2021.
<https://www.rdocumentation.org/packages/mice/versions/3.13.0/topics/mice>.
- Murphy, Kevin P. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press, 2012.
- Olson, Chester L. “Comparative Robustness of Six Tests in Multivariate Analysis of Variance.” *Journal of the American Statistical Association* 69, no. 348 (1974): 894–908. <https://doi.org/10.2307/2286159>.
- Pathak, Manish. “Hierarchical Clustering in R.” DataCamp Community, July 25, 2018.
<https://www.datacamp.com/community/tutorials/hierarchical-clustering-R>.
- Pupale, Rushikesh. “Support Vector Machines(Svm) - an Overview.” Towards Data Science, June 16, 2018.
<https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b428>

00e989#:~:text=SVM%20or%20Support%20Vector%20Machine,separates%20the%20data%20into%20classes.

Rubin, Donald B. “Inference and Missing Data.” *Biometrika* 63, no. 3 (1976): 581–92. <https://doi.org/10.2307/2335739>.

Stahmann, Robert F. “Predicting Graduation Major Field from Freshman Entrance Data.” American Psychological Association. American Psychological Association, 1969. <https://psycnet.apa.org/record/1969-10376-001>.

Starkweather, Jon, and Amanda Kay Moske. “Multinomial Logistic Regression.” University of North Texas, August 2011.

University of Alberta. “Cluster Analysis.” University of Alberta, n.d.

Vink, Gerko, Laurence E. Frank, Jeroen Pannekoek, and Stef van Buuren. “Predictive Mean Matching Imputation of Semicontinuous Variables.” *Statistica Neerlandica* 68, no. 1 (2014): 61–90. <https://doi.org/10.1111/stan.12023>.

Yale-NUS College. “Academic Experience.” Yale-NUS. Accessed April 1, 2022. <https://www.yale-nus.edu.sg/academics/overview/academic-experience/>.

Yale-NUS College. “Academic Regulations,” November 30, 2021.

Yiu, Tony. “Understanding Random Forest.” Towards Data Science, June 12, 2019. <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>.

Appendix A:

A.1 Code

The code can be accessed here:

<https://github.com/napatsakulsaenggrapha/Capstone>

A.2 Literature Review

While there is a multitude of data potentially related to major choice collected at various colleges, little systematic work has been done on predicting a student's actual major choice. One study that has aimed to predict college major, in a similar manner with this capstone, was carried out by Robert Stahmann from the University of Iowa (Stahmann, 1969). Stahmann compared the predictive validity of 3 types of data as predictors of major fields of study. The sample he worked with were bachelor degree graduates from the University of Utah from the class of 1962-1966. The data are: academic achievement test scores, occupational interest inventory scores, and self-expressions of major fields.

1. The occupational interest inventory data contained variables such as fields of interest - personal-social, natural, mechanical, arts, business, and the sciences; types of interests - verbal, manipulative, computational - and the level of interests.
2. The academic achievement test data include scores from various tests such as: cooperative English, mechanics of expressions, vocabulary, total reading, natural science, cooperative mathematics, and etc.
3. The self-expressions of major fields included the response for the questions:

- a. “In what division do you expect to register?” (Arts, Business, Education, Engineering, and etc.)
- b. “In what subjects do you plan to major?”

Each of the 3 datasets were considered systems of data and analyzed separately. Further, Stahmann analyzed males and females separately. Hence, there were 6 combinations of analyses in total. The classification procedure used for the occupational interest and academic achievement data was the multiple discriminant analysis. “Multiple discriminant analysis is a statistical method of combining test scores or other data so as to maximize the differences between the groups and minimize the differences within each group” (Dunn, 1959). On the other hand, the classification for the self-expression data is “tallied for the total sample ... in each major field of study yielding self-prediction information (Stahmann, 1969).”

Stahmann found that self-expressed choice of major fields and intended division of registration was the most accurate predictor of major fields for women; occupational interest data and academic achievement data were the 2nd and 3rd best predictors respectively. The predictions for men, by all data types, were less efficient than that of women. For men, the efficiency of the predictors rank as follows: intended division of registration, first choice of major subject, occupational interest, and academic achievement.

The data set Stahmann used and the one available for this project are similar. The self-expressed choice of major & intended division of registration parallels Yale-NUS’s academic interest fields. The academic achievement data can be seen as an equivalent of the Common curriculum grade data. While Stahmann used his 3 data systems as separate

predictors, this capstone creates models that best predict major choice, using the academic interest and Common Curriculum grade data both jointly and separately.

Another study on the topic of major prediction is that of Allen and Robbins'. While Allen and Robbins did not directly predict students' major choice, they built a model that captured major persistence, defined as a binary variable representing whether "students ... remained in their entering major group into their third year or switched major groups at some time before or during the third year" (Allen and Robbins, 2007). The equivalence of this study in this capstone would be whether or not academic interests persisted into major selection at Yale-NUS. This study utilized a large sample of 48,232 students from 25 four-year institutions. The method used by Allen and Robbins is a two-level hierarchical logistic regression model; the two-level aspect of the model stems from the fact that students were nested within institutions.

A major takeaway from Allen and Robbins' study is that first-year GPA and interest-major composite are key predictors for major persistence. The interest-major composite is based on a student's entering major and two work task curriculums (vocational interests). The estimated regression coefficients were 0.383 for the interest-major composite and 0.360 for the first-year GPA; both these values are statistically significant in predicting major persistence (Allen and Robbins, 2007). While the difference may be small, interest-major composite is a better predictor than first-year GPA. This finding supports what Stahmann found in his study, where intended division of registration and first choice of major subject were better predictors of major than academic achievement score tests.

A.3 Supplementary and Additional Figures

Full Major Name	Acronym
Anthropology	ANTH
Arts and Humanities	ARTS HUM
Economics	ECON
Environmental Studies	ES
Global Affairs	GA
History	HIST
Life Sciences	LS
Literature	LIT
Mathematical, Computational and Statistical Sciences	MCS
Philosophy	PHIL
Physical Sciences	PS
Philosophy, Politics and Economics	PPE
Psychology	PSY
Urban Studies	US

Figure A.1: Major Acronyms

Full Major Name	Acronym
Comparative Social Inquiry	QR
Literature and Humanities 1	LH1
Literature and Humanities 2	LH2
Modern Social Thought	MST
Philosophy and Political Thought 1	PPT1
Philosophy and Political Thought 2	PPT2
Quantitative Reasoning	QR
Scientific Inquiry 1	SI1
Scientific Inquiry 2	SI2

Figure A.2: Common Curriculum Acronym

	PIN	Admit.Year	Academic.Interest.1	Academic.Interest.2	Major	LH1	LH2	PPT1	PPT2	CSI	MST	QR	Science
1	N313106412255	2014	PPE	NA	MCS	0	5.0	0	5.0	0	5.0	4.5	2.833333
2	N359107401288	2014	LS	NA	LS	0	4.0	0	4.0	0	4.5	4.5	3.375000
3	N308110412274	2014	PPE	PSY	LS	0	4.5	0	4.5	0	4.0	4.5	2.833333
4	N337110408274	2014	PPE	ECON	PSY	0	4.0	0	4.5	0	4.0	5.0	3.166667
5	N386110425292	2014	HIST	ANTH	PPE	0	4.5	0	4.5	0	4.5	3.5	2.833333
6	N374110423297	2014	PSY	ARTS HUM	ES	0	4.5	0	4.0	0	4.5	4.0	2.666667
7	N379112424217	2014	PPE	NA	PPE	0	4.5	0	4.5	0	5.0	4.0	2.666667

Figure A.3: Top 7 Rows of the Pre-Imputed Data⁶

	PIN	Admit.Year	AI1	AI2	Major	LH1	LH2	PPT1	PPT2	CSI	MST	QR	Science
1	N313106412255	2014	PPE	NA	MCS	4.5	5.0	5.5	5.0	4.5	5.0	4.5	2.833333
2	N359107401288	2014	LS	NA	LS	3.5	4.0	4.5	4.0	3.0	4.5	4.5	3.375000
3	N308110412274	2014	PPE	PSY	LS	4.5	4.5	5.0	4.5	5.0	4.0	4.5	2.833333
4	N337110408274	2014	PPE	ECON	PSY	4.0	4.0	3.5	4.5	4.5	4.0	5.0	3.166667
5	N386110425292	2014	HIST	ANTH	PPE	4.0	4.5	5.0	4.5	4.0	4.5	3.5	2.833333
6	N374110423297	2014	PSY	ARTS HUM	ES	3.5	4.5	3.5	4.0	3.5	4.5	4.0	2.666667
7	N379112424217	2014	PPE	NA	PPE	4.5	4.5	3.0	4.5	4.0	5.0	4.0	2.666667

Figure A.4: Top 7 Rows of the Data

⁶ Note that the zeroes in the LH1, PPT1, and CSI Columns are there as a placeholders for the to-be imputed grades

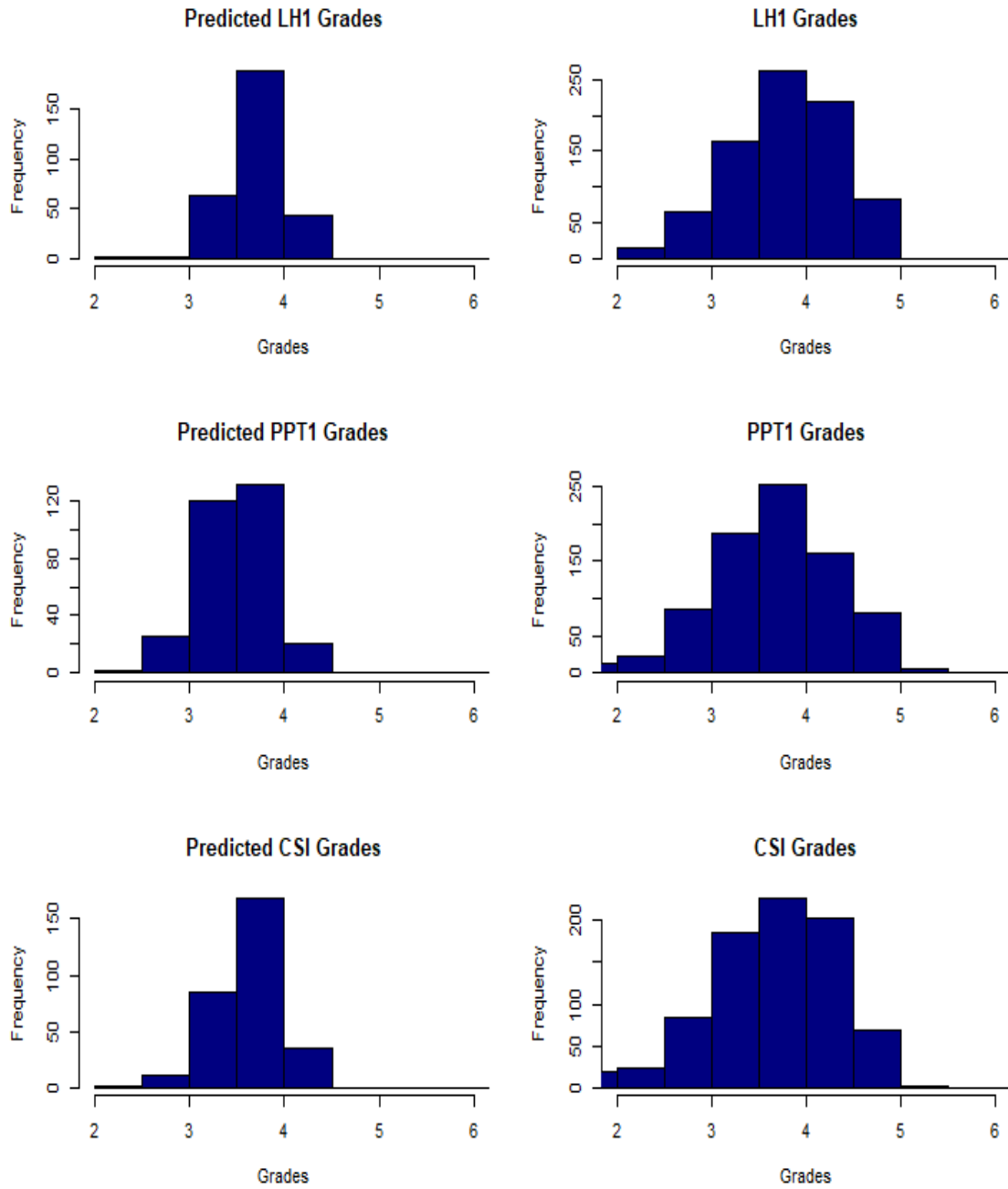


Figure A.5: Distribution of Imputed Data vs. Actual Data (MMR)

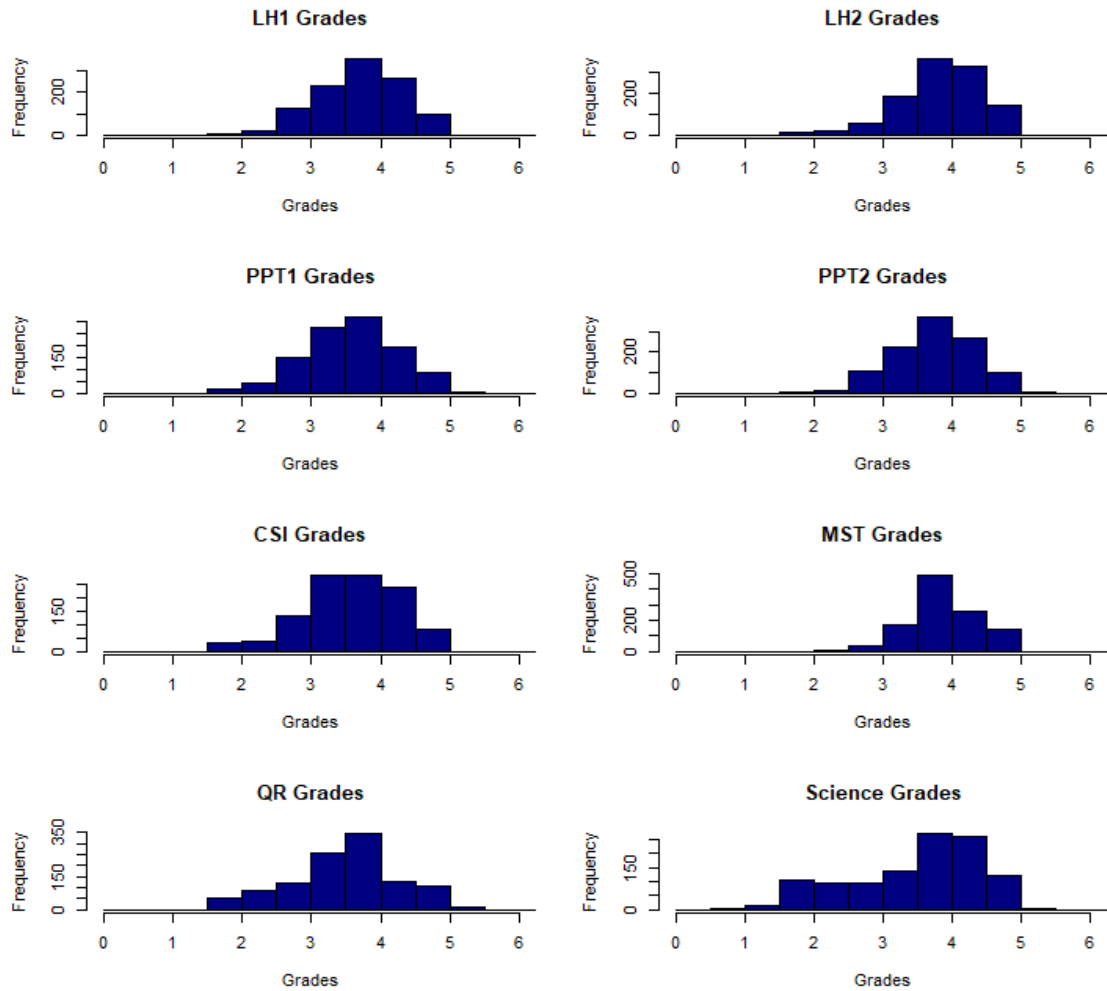


Figure A.6: Common Curriculum Grade Distributions

	PPE	ES	ARTS HUM	GA	HIST	PHIL	ANTH	ECON	US	LIT	PSY	MCS	PS	LS
PPE	52	10	2	26	9	10	10	21	10	2	13	16	0	1
ES	2	31	1	3	0	0	1	3	1	1	0	3	4	2
ARTS HUM	5	5	21	12	3	1	7	2	9	7	4	5	0	2
GA	27	8	1	22	8	4	10	1	20	7	3	5	4	2
HIST	5	1	2	4	15	2	2	1	1	4	1	0	1	2
PHIL	2	1	3	0	0	4	2	0	2	1	2	2	0	0
ANTH	7	9	4	1	4	2	14	1	9	3	3	2	1	2
ECON	18	5	1	4	2	3	3	47	8	2	3	28	0	1
US	0	4	3	1	0	0	2	5	13	1	2	2	0	0
LIT	10	3	5	5	3	2	4	1	4	28	3	2	0	0
PSY	7	9	10	5	3	2	8	1	2	0	50	6	0	4
MCS	3	1	0	2	0	1	0	10	2	2	2	49	4	0
PS	1	5	0	1	0	1	1	3	0	0	1	18	22	6
LS	7	4	0	2	3	0	7	4	3	1	7	6	5	28

Figure A.7: 14 x 14 Adjacency Matrix for All (row) to Major (column) Transitions

	ANTH	ARTS	HUM	ECON	ES	GA	HIST	LIT	LS	MCS	PHIL	PPE	PS	PSY	US
ANTH	0		1	0	1	0	1	1	1	0	1	5	0	1	2
ARTS HUM	1		2	1	0	2	1	1	0	2	0	1	0	2	2
ECON	0		1	13	1	0	0	0	0	1	0	8	2	0	2
ES	1		1	3	5	2	0	1	1	1	1	3	1	3	1
GA	1		0	1	0	5	1	1	0	1	0	5	0	1	3
HIST	1		0	0	0	0	6	2	0	0	0	2	0	1	1
LIT	0		0	0	0	2	2	6	1	1	0	1	0	0	0
LS	0		2	0	0	0	0	0	7	1	1	0	0	3	0
MCS	1		2	7	1	0	0	2	1	19	2	3	0	4	0
PHIL	0		0	1	0	1	0	1	0	2	0	1	0	1	0
PPE	1		1	4	0	3	3	2	0	1	4	11	1	2	1
PS	0		0	0	1	1	1	0	0	3	0	0	1	0	0
PSY	0		1	0	0	1	0	2	3	3	0	3	0	16	1
US	1		2	3	0	1	2	0	1	1	0	1	0	0	5

Figure A.8: Confusion Matrix of Multinomial Logistic Prediction vs. Actual Major

	predict_major														
	ANTH	ARTS	HUM	ECON	ES	GA	HIST	LIT	LS	MCS	PHIL	PPE	PS	PSY	US
ANTH	1		2	0	2	1	1	1	1	0	1	3	0	0	1
ARTS HUM	2		5	2	1	1	1	0	0	0	0	1	0	1	1
ECON	3		0	7	1	0	0	0	2	6	0	6	0	0	3
ES	1		0	3	6	1	0	0	2	3	1	3	0	2	2
GA	0		0	0	1	3	4	0	0	1	1	5	0	1	3
HIST	0		0	0	0	1	6	1	0	0	0	2	0	2	1
LIT	2		0	1	0	1	2	3	0	0	0	4	0	0	0
LS	0		0	0	3	1	0	0	3	0	0	2	1	3	1
MCS	0		1	4	0	1	0	1	1	15	1	7	4	5	2
PHIL	0		0	0	1	1	0	1	0	2	0	1	0	0	1
PPE	1		0	3	0	2	2	2	0	0	3	14	1	4	2
PS	0		0	0	1	1	1	0	0	2	0	0	2	0	0
PSY	3		0	1	1	1	0	1	3	2	1	1	0	14	2
US	1		2	2	0	1	1	0	1	1	0	2	0	0	6

Figure A.9: Confusion Matrix of Random Forest Prediction vs. Actual Major

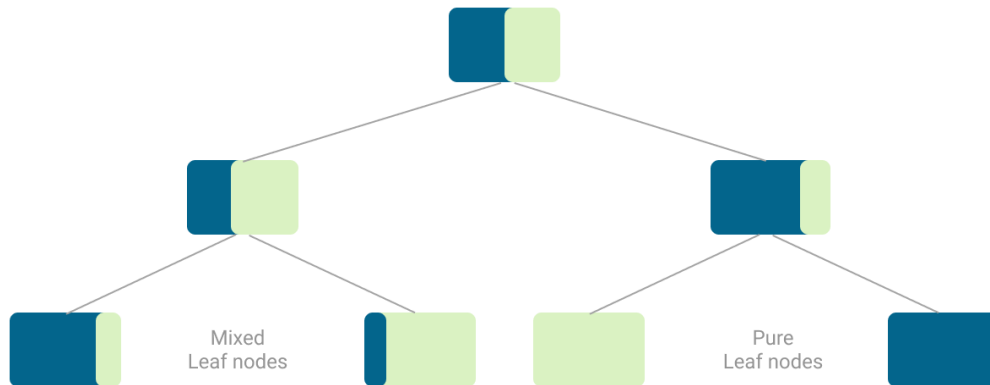
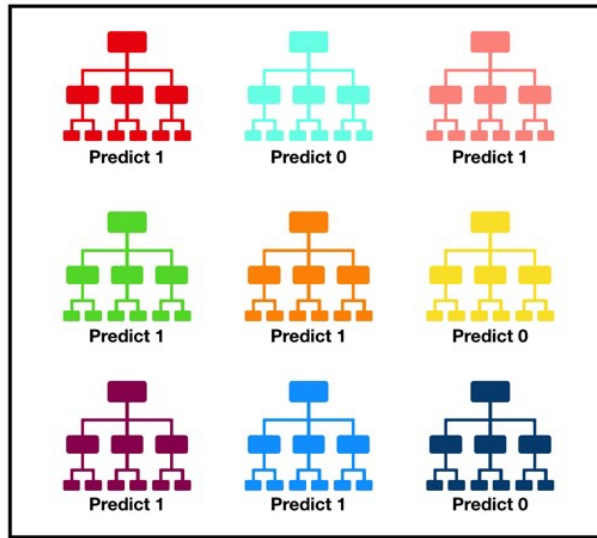


Figure A.10: Example of Mixed and Pure Leaf Nodes (Bento, 2021)



Tally: Six 1s and Three 0s
Prediction: 1

Figure A.11: Example of Random Forest Voting (Yiu, 2019)

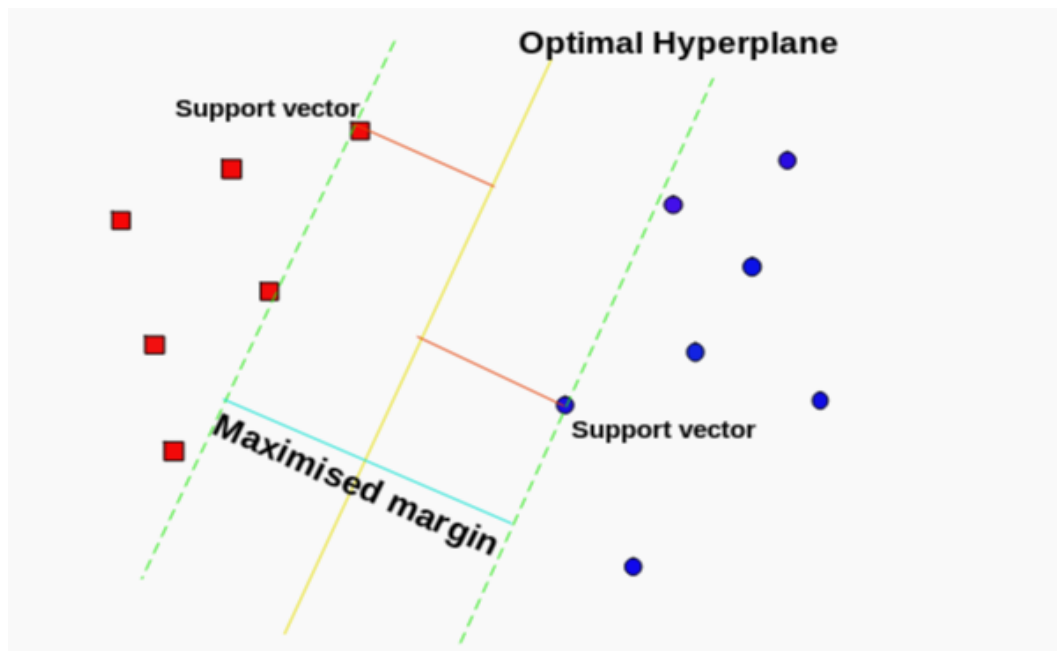


Figure A.12: SVM 2-D Example (Pupale, 2018)

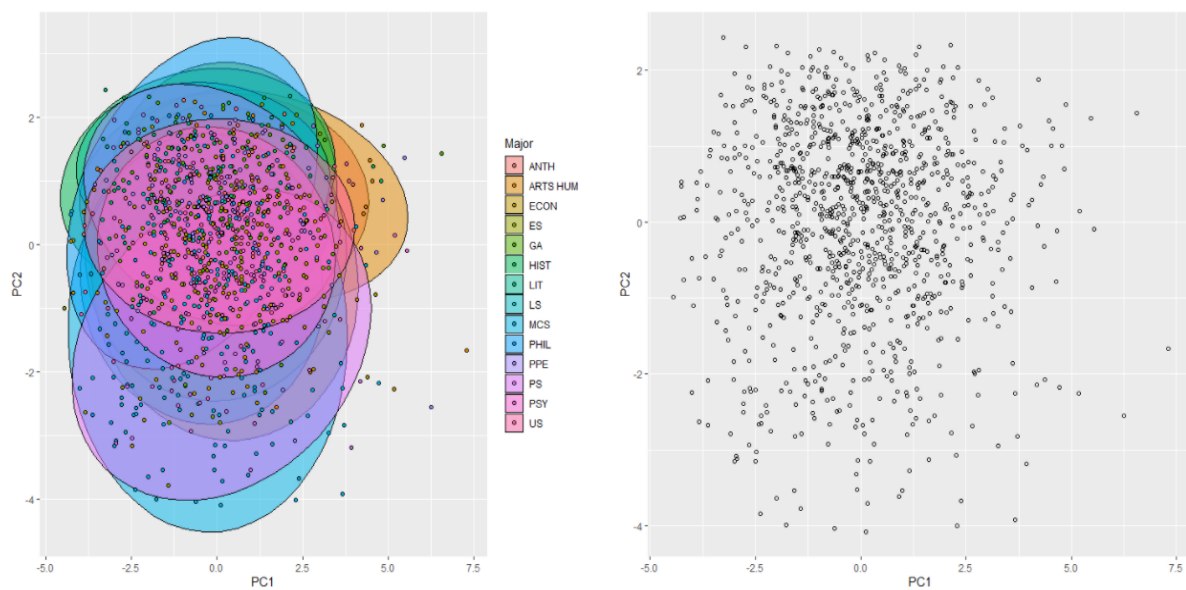


Figure A.13: PCA Results with Labels and No Labels

A.4 R Packages Used For Certain Functions and Algorithms

- Missingness Data Pattern can be created using the function `aggr()` from the `VIM` package
- Circular Transition Diagram can be created using the `chordDiagram()` function from the `circlize` package
- Logistic regression was fitted by using the `glm()` function in R
- The multinomial logistic regression can be implemented simply by training a model using the function `multinom()` from the `nnet` package in R.
- Decision trees can be easily implemented in R by using the `rpart()` function from the `rpart` package.
- The Random Forest Classifier can be implemented in R using the function `randomForest` from the `randomForest` package.
- Gradient Boosting can be implemented in R using the function `gbm()` from the `gbm` package
- AdaBoost can be implemented by using the `boosting()` function from the `adabag` package.
- SVM can be implemented in R by using the function `svm()` from the package `e1071`.
- K-means can be implemented using the `kmeans()` function which is in-built in R
- PCA can be implemented with the function `prcomp()`, built-in in R
- Automatically cutting the dendrogram can be done by the function `cutreeDynamic()` from the package `dynamicTreeCut`