

Utah State University

DigitalCommons@USU

All Graduate Theses and Dissertations

Graduate Studies

5-2021

Data-Driven Recommendation of Academic Options Based on Personality Traits

Aashish Ghimire
Utah State University

Follow this and additional works at: <https://digitalcommons.usu.edu/etd>



Part of the [Databases and Information Systems Commons](#)

Recommended Citation

Ghimire, Aashish, "Data-Driven Recommendation of Academic Options Based on Personality Traits" (2021). *All Graduate Theses and Dissertations*. 8022.

<https://digitalcommons.usu.edu/etd/8022>

This Thesis is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Theses and Dissertations by an authorized administrator of DigitalCommons@USU. For more information, please contact digitalcommons@usu.edu.



DATA-DRIVEN RECOMMENDATION OF ACADEMIC OPTIONS BASED ON
PERSONALITY TRAITS

by

Aashish Ghimire

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Computer Science

Approved:

John Edwards, Ph.D.
Major Professor

Vicki Allan, Ph.D.
Committee Member

Mahdi Nasrullah Al-Ameen, Ph.D.
Committee Member

D. Richard Cutler, Ph.D.
Vice Provost of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2021

Copyright © Aashish Ghimire 2021

All Rights Reserved

ABSTRACT

Data-Driven Recommendation of Academic Options Based on Personality Traits

by

Aashish Ghimire, Master of Science

Utah State University, 2021

Major Professor: John Edwards, Ph.D.
Department: Department of Computer Science

The choice of academic major and, subsequently, an academic institution has a massive effect on a person's career. About 40% of students either transfer to a different major or different college or totally drop out of college within six years because they realize they don't like their academic situation. Various social science research has shown that personality traits play a significant role in academic preference. Still, there has not been a comprehensive, data-driven approach to translate this into academic choice. . There was no existing data for this kind of research, so we surveyed over 500 students. We conducted a survey to capture students' personality traits and preference of college major and used that information to train a machine learning model to predict college major preference. This research validates the viability of using personality traits as indicators for educational preference. We demonstrate that using decision tree, very accurate classification can be done with over 90% accuracy and help with career guidance. Furthermore, we explored the two methods of dimension reduction - one using Principal Component Analysis (PCA) and another relying on Social Science research and using Big-Five personality Traits (also known as OCEAN indexes) to simplify the problem further. With these techniques, the dimension was reduced by half without decreasing the accuracy of our classifier.

With this research, a readily deployable recommendation system is created that can help students find their most enjoyable academic path and aid guidance counselor and parents with their recommendations.

(82 pages)

PUBLIC ABSTRACT

Data-Driven Recommendation of Academic Options Based on Personality Traits

Aashish Ghimire

The choice of academic major and, subsequently, an academic institution has a massive effect on a person's career. It not only determines their career path but their earning potential, professional happiness, etc. [1] About 40% of people who are admitted to a college do not graduate within six years. Yet, very limited resources are available for students to help make those decisions, and each guidance counselor is responsible for roughly 400 to 900 students across the United States. A tool to help these decisions would benefit students, parents, and guidance counselors.

Various research studies have shown that personality traits affect college choice, but there were no studies or tools to utilize this information. With this research, we validate that the personality traits can be used to classify majors, and subsequently, to recommend college majors to students. We identified a method to make that recommendation with more than 90% accuracy. We also analyzed methods of simplifying the personality traits dimension and identified two techniques that can reduce the input data by half and still maintain over 90% accuracy.

To my family, friends, my advisors, professors, mentors, and peers

ACKNOWLEDGMENTS

The biggest acknowledgement to my parents Thakur and Krishna for always being there for me and my siblings Ashim and Asmita for supporting me throughout. I would also like to thank Dr. John Edwards for being an awesome mentor for me throughout as well as committee members Dr. Vicki Allan and Mahdi Nasrullah Al-Ameen. A huge share of credit goes to my friends in Salt Lake City and Logan for making this time memorable. Constant support of Nitesh and Sumu Rijal and Aadarsha Basnet has been huge confidence boost for me. Special thanks to Ashesh, Santosh, Swastik, Bhaskar, Dipen, Ijan, Abiral, Bishav, Hrishiv and Bishal. Similarly, the help and support of staffs Cora, Kaitlyn, Genie and Vicki at the CS Department has been immense.

Aashish Ghimire

CONTENTS

	Page
ABSTRACT	iii
PUBLIC ABSTRACT	v
ACKNOWLEDGMENTS	vii
LIST OF TABLES	x
LIST OF FIGURES	xi
ACRONYMS	xiii
1 INTRODUCTION	1
1.1 Background	1
1.2 Problem today	2
1.3 Scope of Research	4
1.3.1 R1: How effective is the use of personality traits in predicting a preferred college major?	4
1.3.2 R2: How do expert-derived personality traits compare to a data-driven dimension reduction technique ?	4
1.3.3 R3: What is the most effective classification technique to determine college major from personality traits?	4
2 RELATED WORKS	7
2.1 Study of linkage between personality type and academic majors	7
2.2 Use of the personality traits for academic majors and college recommendations:	9
3 METHODOLOGY	10
3.1 Ten-Item Personality Inventory-(TIPI) - 10 Questions for the Personality Type Classification	11
3.2 Big Five Personalty Traits(OCEAN) for dimension reduction	12
3.3 Scoring of survey questions:	12
3.4 Calculations of the mean OCEAN scores	14
3.5 Academic Major Preference	15
3.6 Survey Platform and Data Collection	17
3.6.1 Survey of students at Utah State University	18
3.6.2 survey using Amazon Mechanical Turk (MTurk) platform	18
4 SURVEY DATA	20
4.1 Survey Participation: Amazon Mechanical Turk	20
4.1.1 Gender distribution:	20
4.1.2 Racial distribution:	20

4.1.3	Personality Type questions:	21
4.2	Survey Participation : Utah State University	23
4.2.1	Gender distribution:	23
4.2.2	Racial distribution:	23
4.2.3	Personality Type questions:	23
4.2.4	Sampling Bias and separate processing of data	23
5	CLASSIFICATION	27
5.1	Classification based on raw TIPI survey	27
5.1.1	Features and Target class	28
5.1.2	Tree depth and classification	28
5.2	Dimension Redaction - by OCEAN Indexes	28
5.2.1	Personality Type calculations	30
5.2.2	Features and Target class	30
5.2.3	Tree depth and classification	30
5.2.4	College Majors for different personality traits	31
5.3	Dimension Reduction by Principal Component Analysis (PCA)	36
5.3.1	Features and Target class	36
5.3.2	Tree depth and classification	36
5.4	Accuracy among Raw, OCEAN Dimension Reduced and PCA Dimension Reduced Features	37
5.5	Classification with Neural Network	37
5.6	Classification on Utah State University Survey Data	39
6	DISCUSSION AND CONCLUSION	41
6.1	Key Contributions	41
6.1.1	Viability of use of the personality traits for the data-driven academic recommendation	41
6.1.2	Effect of dimension reduction technique in the recommendations	42
6.1.3	Identification of usable classification technique for recommendation	43
6.2	Future Works	43
	REFERENCES	45
	APPENDICES	47
A	Survey Questionnaire	48
B	IRB Exemption Approval	55
C	IRB Letter of Information	57
D	Distribution of Personality type questions - Amazon Mechanical Turk Survey	59
E	Distribution of Major across Personality type questions - Utah State Univer- sity Survey	65

LIST OF TABLES

Table	Page
2.1 Table showing Holland’s classification of academic majors [2].	8
3.1 Connotation associated with being in different personality traits score	13
3.2 Table showing the number of majors in each general major category	16
4.1 Gender distribution of participant in MTurk and USU data	20
4.2 Racial distribution of participant in Mturk and USU data	21
4.3 MTurk Data: Count of each answers in personality type questions	22
4.4 USU data: Count of each answers in personality type questions	25
5.1 Features and target class for classification with raw TIPI survey responses	28
5.2 The aggregate statistical details of calculated OCEAN Indices - MTurk Survey	30
5.3 The aggregate statistical details of calculated OCEAN Indices - Utah State University Survey	30
5.4 Features and target class for classification with raw data	32
5.5 Features and target class for classification with raw data	36
5.6 Parameters for neural network MLPClassifier	38
5.7 Features and target class for Neural Networks Classification	39
5.8 The result of T-Test between the personality traits derived from two dataset	39

LIST OF FIGURES

Figure	Page
1.1 The student to counselor ration in the different US States [3].	3
3.1 Example of asking one of 10 TIPI questions	16
3.2 Example of question asking user for their preferred major	17
4.1 Answer distribution for the question # 1 above	21
4.2 Answer distribution for the question # 2 above	24
4.3 Answer distribution for the question # 3 above	24
5.1 Accuracy of the decision tree over different tree depth using different features	29
5.2 Distribution of college majors based on Openness using results from the MTurk Survey	31
5.3 Distribution of college majors based on Conscientiousness using results from the MTurk Survey	32
5.4 Distribution of college majors based on Extraversion using results from the MTurk Survey	33
5.5 Distribution of college majors based on Agreeableness using results from the MTurk Survey	34
5.6 Distribution of college majors based on Neuroticism using results from the MTurk Survey	35
5.7 Accuracy of OCEAN and PCA technique between depth 5-20	37
5.8 Accuracy of the different decision tree over different tree depth	38
5.9 Accuracy of the different decision tree over different tree depth	40
D.1 Answer distribution for the question Q1	59
D.2 Answer distribution for the question Q2	60
D.3 Answer distribution for the question Q3	60

D.4	Answer distribution for the question Q4	61
D.5	Answer distribution for the question Q5	61
D.6	Answer distribution for the question Q6	62
D.7	Answer distribution for the question Q7	62
D.8	Answer distribution for the question Q8	63
D.9	Answer distribution for the question Q9	63
D.10	Answer distribution for the question Q10	64
E.1	Distribution of college majors based on Openness	65
E.2	Distribution of college majors based on Conscientiousness	66
E.3	Distribution of college majors based on Extraversion	67
E.4	Distribution of college majors based on Agreeableness	68
E.5	Distribution of college majors based on Neuroticism	69

ACRONYMS

TIPI	Ten-Item Personality Inventory
DT	Decision Tree
OCEAN	Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism
MTurk	Amazon Mechanical Turk
USU	Utah State University
IRB	Institutional Review Board
SAT	Scholastic Assessment Test
GPA	Grade Point Average
PCA	Principal Component Analysis
QR	Quick Response
LBFGS	Linear Broyden Fletcher Goldfarb Shanno Algorithm
DNN	Deep Neural Networks

CHAPTER 1

INTRODUCTION

1.1 Background

The choices of academic major and institution are among the most fundamental steps in a person's career. A student has to make that decision early in their career and often without sufficient information. According to the department of education, the overall 6-year graduation rate for first-time, full-time undergraduate students who began seeking a bachelor's degree at 4-year degree-granting institutions in fall 2012 was 62 percent. That is, by 2018, some 62 percent of students had completed a bachelor's degree at the same institution where they started in 2012. The 6-year graduation rate was 61 percent at public institutions, 67 percent at private nonprofit institutions, and 25 percent at private for-profit institutions [4]. This 6-year limit does not take the breaks taken for religious mission, military services etc. For the cohort starting in 2011, the eight-year graduation rate was only 61.8 % [5].

Six-year graduation rates for first-time, full-time undergraduate students who began seeking a bachelor's degree at 4-year degree-granting institutions in fall 2012 varied according to institutional selectivity. In particular, 6-year graduation rates were highest at institutions that were the most selective (i.e., those with acceptance rates of less than 25 percent) and were lowest at institutions that were the least selective (i.e., those with an open admissions policy). For example, at 4-year institutions with an open admissions policy, 34 percent of students completed a bachelor's degree within six years. At 4-year institutions with acceptance rates of less than 25 percent, the 6-year graduation rate was 90 percent. Similarly, the graduation rate of the schools where most students live on campus is significantly higher than that of schools where most students commute. [4].

The reason some colleges do so much better than others consists of multiple layers. A

part of the answer involves structure. Students tend to do better when they are following defined academic path rather than randomly signing up for classes without a clear picture of the end goal. It has been shown that schools that provide some sort of degree plan or "pre-majors" have a much better retention rate compared to others. [6]

1.2 Problem today

These data shows the effect that guidance counseling can have on student success, and also shows the importance of choice of institution and major. According to a CNBC report, about 25 percent of all students transfer college at some point before graduation. These statistics clearly demonstrate that the importance of choosing the right campus, academic major, campus setting, etc. makes a lot of difference for a student whether they will graduate. Traditionally, schools provide a guidance counselor to help students making these decisions. According to The National Association for College Admission Counseling (NACAC), the national student-to-counselor ratio is 482:1. In some states like Arizona, there are as many as 924 students per guidance counselor [3]. At this ratio, a guidance counselor cannot realistically suggest the best-fit college or major for each student.

In order to aid these students with making those decisions, guidance counselors could potentially use different data-driven and well-informed approaches. But, as of now, the state of the art guidance counselor tool is the college database that can be used to filter the institution based on size, location, major, etc. but nothing beyond that according to the market research by the project partner Graphium Inc. There are some commercial services available directly to the students based on their preferences, but they are riddled with university advertisements and sponsored sections, where colleges that pay are boosted higher up on the chart and are not very distinguishable from the real recommendation. Similarly, the College Board - the most popular college recommendation website - is run by the same institution that conducts SAT tests. It puts the full weight on college test scores and GPA but not the the strengths, weaknesses, and personality of the students.

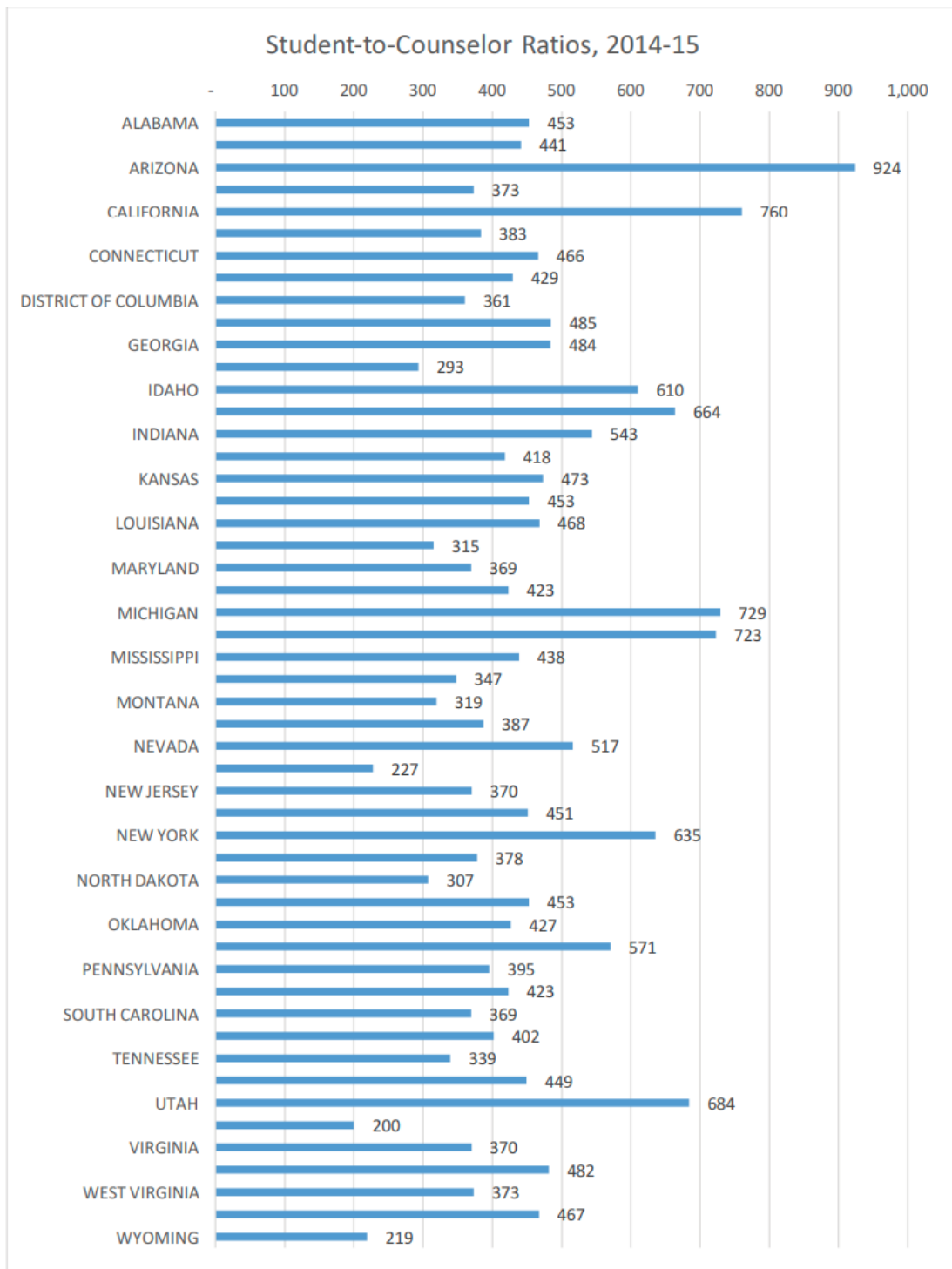


Fig. 1.1: The student to counselor ration in the different US States [3].

1.3 Scope of Research

This research aimed to is a part of a larger recommendation suite launched by our project partner, Graphium Inc, who will be recommending not just college majors, but other areas to focus on as well, including extracurricular activities and so on. The scope of this study comprises three research questions:

1.3.1 R1: How effective is the use of personality traits in predicting a preferred college major?

The use of personality traits for academic majors has not been studied or used in any data-driven way. This study explores the viability of the use of personality traits in academic major recommendations.

1.3.2 R2: How do expert-derived personality traits compare to a data-driven dimension reduction technique ?

Using personality type questionnaires to group a person into a different personality is the social science equivalent of dimension reduction. Prior knowledge of psychology is used to make that educated guess for dimension reduction. This study compares the performance of such a technique with a popular data-driven dimension reduction technique - Principal Component Analysis (PCA).

1.3.3 R3: What is the most effective classification technique to determine college major from personality traits?

There exist many different classification techniques with their own weakness and strength on different kinds of data sets. This research identifies a classification technique that would be most effective to be used as a classifier for the recommendation system that would use the questionnaire as features to make academic major recommendations.

When these questions are answered, the effectiveness of using personality traits in academic decision-making can be determined. This research can lead to the identification of

the effective classification technique and possibly contribute to developing a tool to recommend an academic major based on personality traits. This will make a huge difference for students who do not have adequate access to guidance counselors to help them with these kinds of decisions. This would also provide a good starting point for the career exploration for students as well as a resource for parents and guidance counselors in helping students make their decisions.

For this study, a labeled dataset that has features that measure the personality traits of a student as well as their choice of academic majors was needed. In order to obtain that information, we decided to conduct a survey to obtain the labeled data. While designing the survey, special consideration was made to keep it consistent with the app sign up and onboarding process of the Spotivity app, the career guide software developed by Graphium Inc that this project aims to contribute to. Keeping the survey question consistent with an existing survey in the app will make the integration and the deployment of this feature seamless. For that reason, same demography and personality questionnaires were asked as for the feature variable for the study and academic majors to serve as the labeled target variables. These surveys were launched simultaneously on two platforms - Amazon Mechanical Turk as well as internally in Utah State University.

Once these data were obtained through the surveys, they were cleaned and filtered to exclude any spam and short-duration surveys. Those two surveys were analyzed separately in order to be able to make a comparison later. These filtered data were used to train the classifier to predict academic major preference using three different sets of features. First, the raw input - the answers to 10 personality questions, were used as the feature sets. Similarly, the OCEAN scores were calculated using the published technique of reducing the ten answers into five personality traits of openness, conscientiousness, extraversion, agreeableness, and neuroticism. Finally, the principal Component Analysis technique was used to reduce the dimension of the raw data and was used as feature sets as well. The same sets of classification techniques were used for all three feature sets. I used the Decision Tree (DT) classifier across different depths of the tree. The data was split in a 70:30 ratio

for training data and test data to evaluate the performance of our classifier.

After the evaluation of the predicted major against the three choices made by each survey respondent, it can be concluded that the use of personality traits for the recommendation of academic major is viable in real-world use. The test accuracy of over 94 percent was achieved with all three feature sets. Regarding the depth of the decision tree, a tree with 15 levels of depth will yield to about 90 percent, and the gains are minimal beyond the depth of 20, hovering about 94 to 95 percent.

Comparing the accuracy achieved by using the OCEAN Index as the feature sets with the feature sets from the Principal Component Analysis, both techniques perform roughly equally with the tree of depth 20 or more. However, the use of OCEAN scores provides an additional layer of information - an interpretable feature set that has a real-world meaning, unlike PCA features. These scores can be used to give a personalized recommendation for the student to strengthen their skills in certain area.

CHAPTER 2

RELATED WORKS

2.1 Study of linkage between personality type and academic majors

There have been several studies to classify human traits into different personality types and to study student performance in different academic majors. Most famously, Holland [7] classified academic major with six personality types. It is shown in table 2.1. The analysis of covariance results indicated that four of the five expectation scales were significantly related to students' personality types. In contrast, only two of the expectation scales were significantly related to environment types. This classification was widely used across different papers in psychology research for decades [2]. Allred et al. studied the validity of the stereotypes surrounding the choice of academic majors and stress level in different academic disciplines [8].

Giacomino and Akers in 1998 conducted a study within the business school. They found out that the choice of a specific major is linked to personality traits, values, and interpersonal behavior and is again mediated by gender differences [9]. In 2010, Pringle et al. studied personality type and their role in the academic major, but this was limited to the Business school exploring different sub-branches within the School of Business [10]. In 2016, Eide et al. conducted a comprehensive study that reviewed 11 different papers that studied the relationship between personality types and academic major [1]. These studies use newer "Big Five Personality Traits" - openness, conscientiousness, extraversion, agreeableness, and neuroticism . The initial model was advanced by Ernest Tupes and Raymond Christal in 1961 but failed to reach an academic audience until the 1980s [11]. In 1990, J.M. Digman advanced his five-factor model of personality [12], which Lewis Goldberg extended to the highest level of organization [13]. These five overarching domains have been found to contain and subsume most known personality traits and are assumed to represent

Investigative	Artistic	Social	enterprising
General Biology Biochemistry Biophysics Botany Marine (Life) Science Microbiology Bacteriology Zoology Finance Aeronautical Engineering Astronautical Engineering Civil Engineering Chemical Engineering Astronomy Atmospheric Science Chemistry Earth Science Mathematics Physics Statistics Pharmacy Premedical Predental Preveterinary Anthropology Economics Ethnic Studies Geography Sociology	Arts English Language Literature Music Speech Theatre Drama Music Art Education Architecture	History Philosophy Theology Religion Elementary Education Physical Education Recreation Special Education Home Economics Library Science Nursing Political Science Psychology Social Work Women's Studies Law Enforcement	Journalism Business Administration Marketing Management Business Education Industrial Engineering Communications Computer Science
Realistic	Conventional		
Electrical Engineering Mechanical Engineering Marine Science Drafting/Design Military Science	Accounting Secretarial Studies Data Processing		

Table 2.1: Table showing Holland's classification of academic majors [2].

the basic structure behind all personality traits. In this study, these five personality traits will be used for our recommendation.

2.2 Use of the personality traits for academic majors and college recommendations:

In 1982, Cebula conducted a study to examine what motivates a student to choose a certain academic focus [14]. However, this was limited to their Business school, and the study indicated that the potential of future income was the biggest driver. Zheng et al. conducted another study in 2002 to examine the predictor for academic success for freshmen and found some aspects of personality traits, like the self-perception of abilities, playing a major role [15].

Most prior works are from a social science perspective and not taken as a classification problem. In my research, I explore the viability of using personality traits data to make recommendations for academic majors.

CHAPTER 3

METHODOLOGY

This study was designed to answer three research questions: How effective is the use of personality traits in predicting a preferred college major; How do expert-derived personality traits compare to a data-driven dimension reduction technique; and What is the most effective classification technique to determine a college major from personality traits? Since there isn't available data to be used for the classification, I, as a major part of this study, collect the data in the form of a survey. This survey asks ten probing personality questions, as described in section 3.2. It also asks three questions related to college major preference, along with a couple of demographics questions. With these survey responses, the scores for five personality traits (using a psychology expert-derived dimension reduction) are also calculated from the answers to the ten personality questions and are used for the classifications. For comparison, I also build classification models using the answers to ten questions and from the components derived with Principal Component Analysis (PCA). I compare the performance of feature sets.

This research is part of a larger project with partner company Graphium Inc., which is building a comprehensive recommendation suite through the Spotivity app that not only covers the college and major recommendation but various other after school activities, scholarships and such. For the scope of my studies, I limit the scope to the academic major recommendation. The questions for the survey were heavily guided by the type of question asked for the user of that platform to keep the consistency and for easier deployment.

There are different sets of questions that can be asked to get information on one's personality. Oregon Research Institute maintains a public domain collection which currently contains 3329 items [16]. There are various versions of the tests, ranging from as few as five items to more than 100 questions for the test. For Example, Goldberg et al. published a 120 question test called IPIP-NEO-120 [13]. It was widely used, and various research has

been done since to reduce and simplify it.

For our study, we needed a survey that is short and simple enough to be completed without burdening users because it would be done online. Gosling et. al. from University of Texas developed the Ten-Item Personality Inventory (TIPI) in 2003 to meet the need for very short measures of the Big Five for time-limited contexts or large survey questionnaires [17]. Even though this survey is relatively short, it has been widely adopted. The ten questions of TIPI were used for this study.

3.1 Ten-Item Personality Inventory-(TIPI) - 10 Questions for the Personality Type Classification

In the user survey, the following ten questions were asked to gauge the personality type of the user:

1. I see myself as extroverted and enthusiastic.
2. I see myself as critical and quarrelsome.
3. I see myself as dependable and self-disciplined.
4. I see myself as anxious and easily upset.
5. I see myself as complex and open to new experiences.
6. I see myself as reserved and quiet.
7. I see myself as sympathetic and warm.
8. I see myself as disorganized and careless.
9. I see myself as calm and emotionally stable.
10. I see myself as conventional and uncreative.

3.2 Big Five Personalty Traits(OCEAN) for dimension reduction

For classifying the user and better understanding their personality type, we use a Big Five Personality Trait index - a method that uses openness, conscientiousness, extraversion, agreeableness, and neuroticism. This is often referred as the OCEAN or CANOE index. In our study, we will assign the user as being either negative (score of 2.5 or under), neutral (between 2.5 to 5.5) or positive (5.5 and over). These three buckets of being negative, positive, and negative in each of five personality traits gives interpretable categories - which can be used to provide personalized help to the students. For example, if someone is in the 'negative' openness bucket, they can be helped with tasks and activities for boosting confidence and public speaking. There are certain connotations based on a personality type being positive or negative. If the user is in the neutral bucket, no inference is made from that index. A brief introduction of those five personality types and the connotation associated with each bucket of those traits are described in the table [3.1] as described by Brick et al. [18].

3.3 Scoring of survey questions:

A seven-level Likert Scale [19] was used to record the responses to the ten questions on the personality of TIPI (see section 3.1) . This is the recommended scoring method by the creator of TIPI [17] . The possible options and their corresponding score was:

1. Strongly disagree - score of 1
2. Disagree - score of 2
3. Somewhat disagree - score of 3
4. Neither agree nor disagree - score of 4
5. Somewhat agree - score of 5
6. Agree - score of 6
7. Strongly agree - score of 7

Personality Traits	Positive (≥ 5.5) Connotations	Negative (≤ 2.5) Connotations
Openness	imagination, training performance, academic achievement, cognitive functioning, creative achievement, independence	training performance (negative), academic achievement (negative), practicality, conformity, negative affect, stress
Conscientiousness	job performance, academic Achievement motivation, judgement, health, stress (negative), problem-solving, goal-orientation, longevity	job performance (negative), academic achievement (negative), impulsivity, carelessness, cognitive decline
Extraversion	teamwork, affection, relationship satisfaction, positive affect, well-being,	teamwork (negative), reservations, somber
Agreeableness	positive relations with others, job performance, academic achievement, softhearted, generosity, altruism, relationship satisfaction, positive emotions	negative relations with others, job performance (negative), academic achievement (negative), uncooperative, skepticism
Neuroticism	teamwork (negative), insecurity, health risk, low well-being in late adulthood	teamwork, satisfaction

Table 3.1: Connotation associated with being in different personality traits score

3.4 Calculations of the mean OCEAN scores

Scores for the OCEAN personality index can be from the ten scores of the TIPI. Scores for each of the five OCEAN categories are calculated using the following formulae as per the Gossling et. al. worksheet [17], where TQ1, TQ2 ... TQ10 refer to answers to the 10 TIPI personality questions:

$$Openness = \frac{TP5 + (8 - TP10)}{2}$$

$$Conscientiousness = \frac{TP1 + (8 - TP8)}{2}$$

$$Extraversion = \frac{TP1 + (8 - TP6)}{2}$$

$$Agreeableness = \frac{(8 - TP2) + TP7}{2}$$

$$Neuroticism = \frac{(8 - TP4) + TP9}{2}$$

Along with these 10 questions for the personality types, our survey contained three demographics questions:

1. Age: This was used mainly to make sure the survey is limited to respondents aged between 18 to 28. We decided on this limitation because this ensures the responses from people who are either in college or thinking about going to college. Additional filters were used in Amazon Mechanical Turk (MTurk) platform, as described in the next section.

2. Gender: Three options were provided as Male, Female and Non-Binary as per suggestion from Institutional Review Board (IRB).
3. Race/Ethnicity: Six options were provided: White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander and Others. This is as per the latest identification options used by the US Census Bureau.
4. Attention checker: MTurk survey can be, at times, susceptible to automated bots taking surveys for the rewards. Besides using MTurk's tools to weed out these survey takers as mentioned in section 2.20.2, an attention checker was built into the survey. After completing the first two sections, a simple arithmetic problem was presented, and respondents were only allowed to proceed and compensated if they passed the attention checker questions. If they answered incorrectly, the survey was instantly ended.

3.5 Academic Major Preference

The list of all college majors listed by at the US Department of education was obtained from their public database [20]. This yielded 397 unique majors. To narrow it down, only majors offered at more than 100 schools were picked, which left 261 majors. From there, all the majors are divided into 14 general categories - as per how these majors are commonly classified in their school's organization structure. These are shown in the table 3.2.

After asking three demographics and ten personality questions, survey takers were provided with an attention checker question. See figure 3.1 for an example of a question regarding personality type. When they pass the attention checker, users were provided the question where they were presented a list of five academic majors randomly selected from the 14 major categories and asked to choose the major that interests them the most. With limited number of training data, recommending individual major is not feasible and recommending a category makes it easier if a student is to switch major within the department compared to totally different one. See figure 3.2 for an example of a question regarding the academic major. To aid the user, three common majors in that category were given

Categories	# of majors included
Management	29
Business	13
Health Service	25
Law / Administration	15
Education	10
Engineering	46
Social and Behavioural Science	21
Life Science	30
Language and Literature	16
Vocational	16
Arts	9
Communication & Media	8
Physical Science	16
Philosophy & Theology	10

Table 3.2: Table showing the number of majors in each general major category

I see myself as sympathetic and warm

- Strongly agree
- Agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Disagree
- Strongly disagree

Fig. 3.1: Example of asking one of 10 TIPI questions

Of the college majors shown below, select the one that interests you the most:

- Language and Literature (Literature, English, Foreign Language, etc.)
- Life Science (Biology, Ecology, Neuroscience, etc)
- Education (Elementary Education, Special Ed and Teaching, Curriculum Development, etc)
- Health Services (Medical Doctor, Nursing, Dental, etc)
- Management (Business Administration, Finance, Management Science, etc)

Fig. 3.2: Example of question asking user for their preferred major

as examples along with the questionnaires. For Example, Physical Science (Physics, Mathematics, Chemistry, etc.). The whole list of examples provided for each general category is listed in Appendix A. For each user, the choices they were offered, and their selections were recorded. These questions were repeated two more times to ensure uniform coverage of all the majors. The decision was made to break this into three questions so that the user read all the options for those questions. Reading through all 14 options for a major for each question could be tedious, and the user might be inclined to pick a random answer. The order and choice of an option was randomized so as to not introduce bias by grouping certain majors together. The options were replaced for each iteration so that if the user's top two or three choices are all grouped in the first question, they will get a chance to select those in the second iteration. There is a small chance of about 2% that a user might not see any of their top three major preferences. This study makes an assumption that there would be a large enough sample size that eventually, it would average out for the top preference major to be shown.

3.6 Survey Platform and Data Collection

The survey was designed in Qualtrics's survey platform. The survey is attached in Appendix A. The survey questions were partly based on the questions that are already being asked on the platform that our project partner is deploying.

Survey responses were collected from two different sources - Utah State University students and Amazon Mechanical Turk workers. For the sake of comparison, these two groups of data were analyzed and used separately. The USU Institutional Review Board (IRB) approved the surveys for both Utah State University Students and Amazon Mechanical Turk (MTurk) platform with a slightly varied questionnaire (to fit the logistics of validating and paying MTurk's respondents (see Appendices [B](#) and [C](#)). Participants were compensated equally - one dollar in Amazon gift cards.

3.6.1 Survey of students at Utah State University

A survey was distributed to students at Utah State University. For the survey recruitment, it was distributed by Quick Response (QR) Code to the cadets of the US Army Reserves Officer Training Corps (ROTC) program at the University. Similarly, an announcement was sent by a Teaching Assistant in the general psychology class at USU. ROTC program and the general education class were chosen to maximize the diversity of survey takers from different academic disciplines. Being a USU student automatically implied the location of the United States and the status of college students. The age was of the respondent was restricted to 25 or under.

3.6.2 survey using Amazon Mechanical Turk (MTurk) platform

The Amazon Mechanical Turk (MTurk) platform was used to gather more survey results. MTurk is a commercial survey and crowd-sourcing platform where a user completes a task for a small reward. In MTurk, this survey was restricted to United States residents. Similarly, the survey was only made available to users with at least a High School diploma using MTurk's premium filter purchase. To avoid the spammy responses for the reward, Amazon provides an option to limit the exposure of the surveys to a set of reliable survey takers. The reliability of survey takers is generally measured in terms of acceptance rate. MTurk allows the survey requester to accept or reject surveys based on the quality of works. For example, if the user have a tendency of completing a survey in a very short time, or use other programmatic tools to complete it, they get rejected more often and have lower

acceptance rate. For the purpose of this study, MTurk filter was used to limit responses to users with more than 95% acceptance rate and 500 accepted surveys.

CHAPTER 4

SURVEY DATA

The survey, as described in Chapter 2, was launched to two platforms simultaneously - on Amazon Mechanical Turk (MTurk) and at Utah State University. The Utah State University Survey was left open for two weeks, with the majority of responses coming on the first three days. For MTurk, the first batch of 20 surveys was launched to validate our setup, and later another 380 surveys were collected in the space of 3 days.

4.1 Survey Participation: Amazon Mechanical Turk

There were a total of 728 responses from Amazon mechanical Turk. However, when we select only those who are aged between 18 to 25 and who fully completed the survey, we have 420 left. We next looked into the completion time and removed any surveys that took less than 60 seconds. After these filter criteria, 355 responses remained. For that valid data, the mean completion time is 142 seconds, and the median is 106 seconds.

4.1.1 Gender distribution:

The MTurk survey participant 4.1 are fairly even in gender distribution as compared to USU data:

Gender	MTurk Count	USU Count
Male	176	36
Female	171	151
Non-Binary	8	1

Table 4.1: Gender distribution of participant in MTurk and USU data

4.1.2 Racial distribution:

The racial distribution of survey is as listed in the table 4.2 below:

Race	MTurk Count	MTurk %	USU Count	USU %
White	266	75%	164	87.23%
Black or African American	44	12.67%	2	1.06%
American Indian or Alaska Native	3	0.8%	3	1.59%
Asian	41	11.54%	15	7.97%
Native Hawaiian or Other Pacific Islander	1	0.28%	1	0.53%
Others	11	3.09%	3	1.59%

Table 4.2: Racial distribution of participant in Mturk and USU data

4.1.3 Personality Type questions:

Of 10 questions in the Ten Item Personality Measure (TIPI), half are reverse-scored. It means the survey is designed in such a way that for half of the question, it is expected that answers are skewed towards "Disagree," and half skewed towards "Agree". A sample plot is attached figure. 4.1 . All other plots are listed in Appendix D. The table 4.3 below shows the count of each response.

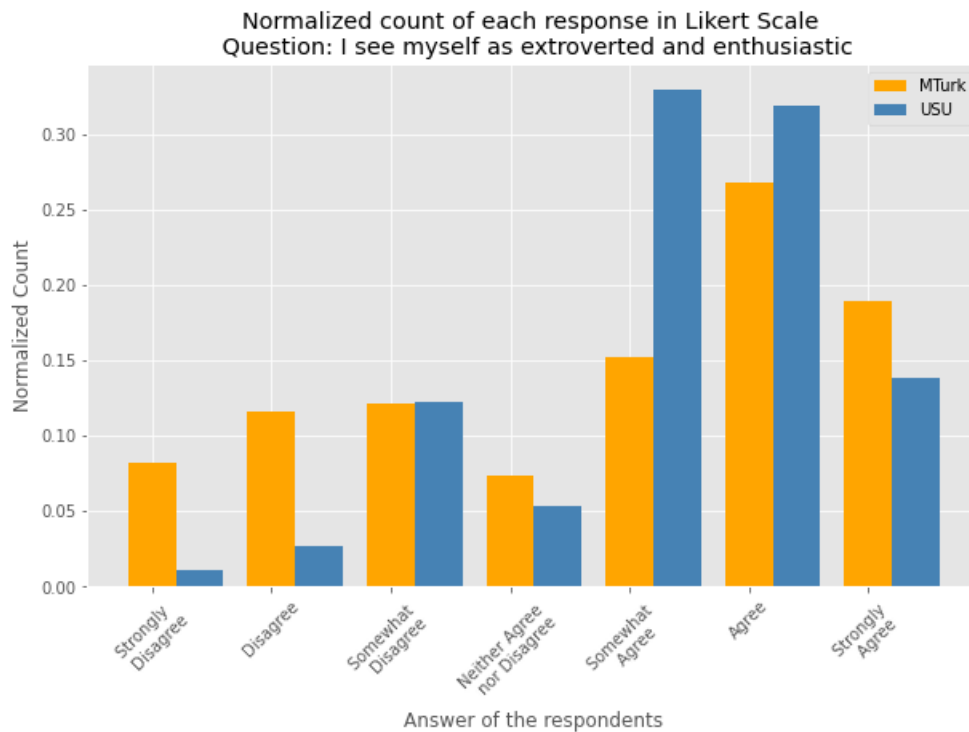


Fig. 4.1: Answer distribution for the question # 1 above

Question / Options	Strongly Agree	Agree	Somewhat Agree	Neither agree nor disagree	Somewhat Disagree	Disagree	Strongly Disagree
I see myself as extroverted and enthusiastic	67	95	54	26	43	41	29
I see myself as critical and quarrelsome	30	59	67	26	46	67	60
I see myself as dependable and self-disciplined	120	116	74	17	21	8	2
I see myself as anxious and easily upset	43	46	71	24	78	71	52
I see myself as open to new experiences and complex	96	102	86	36	21	8	6
I see myself as reserved and quiet	82	100	65	28	28	39	13
I see myself as sympathetic and warm	99	125	63	26	28	9	6
I see myself as disorganized and careless	20	37	47	30	45	85	91
I see myself as calm and emotionally stable	89	121	62	38	30	9	6
I see myself as conventional and uncreative	28	43	34	46	52	79	73

Table 4.3: MTurk Data: Count of each answers in personality type questions

In the next chapter, these survey responses will be used to calculate the personality type and make recommendations.

4.2 Survey Participation : Utah State University

There were a total of 204 responses from Utah State University. However, when we select only those who are aged between 18 to 25 and fully completed the survey, we have 195 left. We next looked into the completion time and removed the surveys that took less than 60 seconds. After these filter criteria, 188 responses remained. For that valid data, the mean completion time is 221 seconds, and the median is 146 seconds.

4.2.1 Gender distribution:

The survey participant is disproportionately skewed towards female participant in the Utah State University data, as shown in the table [4.1](#).

4.2.2 Racial distribution:

The racial distribution of survey is as listed in the table [4.2](#) and discussed in more detail in section [4.2.4](#):

4.2.3 Personality Type questions:

For the ten personality questions (TIPI) to find out the personality type, there is a different distribution depending upon the positive and negative framing of the question. See figure. [4.2](#) and figure. [4.3](#) as examples. All other plots are listed in Appendix [E](#). Table [4.4](#) below shows the count of each response.

4.2.4 Sampling Bias and separate processing of data

As per most research done among two different communities, this survey could be vulnerable to a sampling bias. The US population is about 13.4% Black or African American [[21](#)], but the proportion in Utah State University's survey is under 1%. While this is closer to Utah State's population (1.5%), this is very different from the overall US population. The

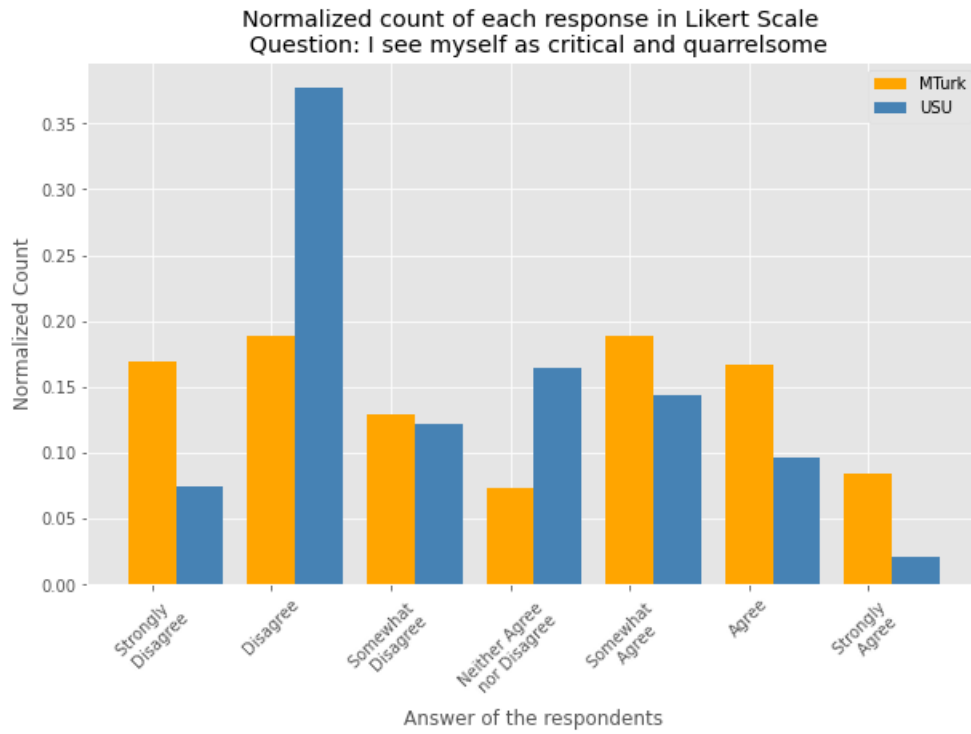


Fig. 4.2: Answer distribution for the question # 2 above

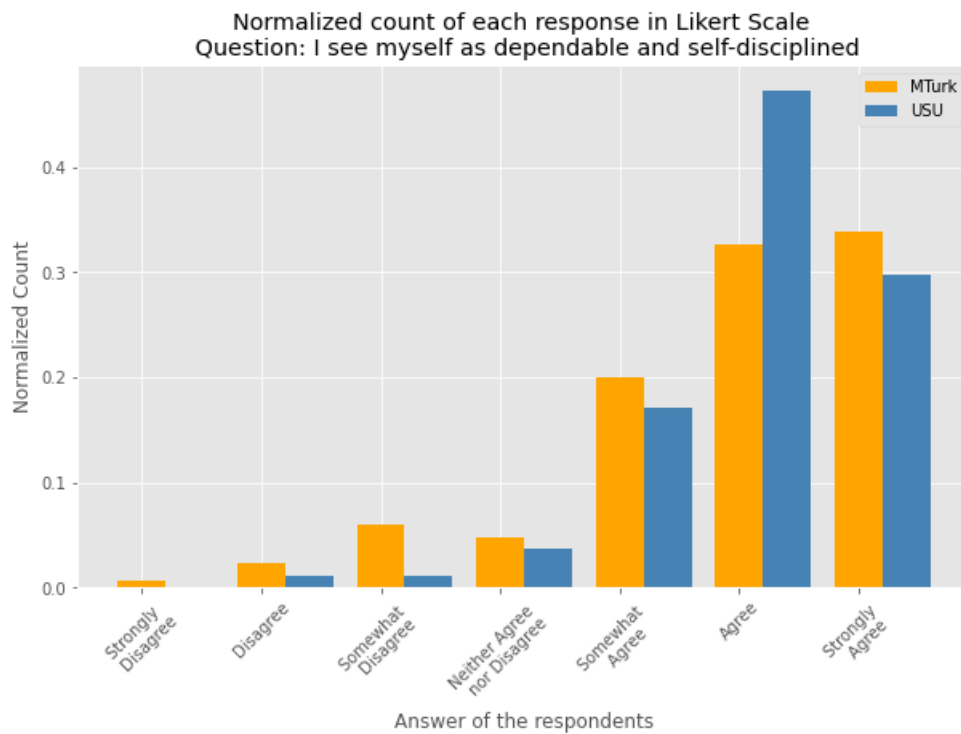


Fig. 4.3: Answer distribution for the question # 3 above

Question / Options	Strongly Agree	Agree	Somewhat Agree	Neither agree nor disagree	Somewhat Disagree	Disagree	Strongly Disagree
I see myself as extroverted and enthusiastic	26	60	63	10	23	5	2
I see myself as critical and quarrelsome	4	18	17	31	23	71	14
I see myself as dependable and self-disciplined	56	89	32	7	2	2	0
I see myself as anxious and easily upset	13	20	46	17	30	52	10
I see myself as open to new experiences and complex	32	79	63	11	3	0	0
I see myself as reserved and quiet	14	25	61	25	26	27	10
I see myself as sympathetic and warm	56	83	39	6	4	0	0
I see myself as disorganized and careless	1	3	24	7	28	62	63
I see myself as calm and emotionally stable	23	73	44	19	26	2	1
I see myself as conventional and uncreative	9	6	40	28	40	43	21

Table 4.4: USU data: Count of each answers in personality type questions

racial composition is MTurk's data is much closer to the US demography. For example, it has about 12.5% of Black or African American population - very similar to national averages.

In terms of gender as well, MTurk's data is very well balanced, while USU's survey is skewed towards higher female participation. In terms of time taken to complete the survey, USU has a much higher time (mean of 221 seconds as opposed to 142 seconds at MTurk). This generally signifies that the user put thought into their answers. Similarly, USU data is guaranteed to be students, but the same could not be said for the MTurk survey.

Therefore, these data are analyzed separately. This not only helps to see the performance of our classifier on a different independent dataset but also helps to identify if skew in one dataset is causing issues.

CHAPTER 5

CLASSIFICATION

The three primary research questions for these projects are (a) R1: How effective is the use of personality traits in predicting a preferred college major? (b) R2: How do expert-derived personality traits compare to a data-driven dimension reduction technique? and (c) R3: What is the most effective classification technique to determine a college major from personality traits? I launched the survey, collected the data, and did some exploration of data as described in the first four chapters.

This all leads to the method of classification which will be used to make the academic major recommendations. The accuracy of the test dataset validates our hypothesis that personality traits are suitable features for major classification. Similarly, the comparison between the performance of the OCEAN feature set (which is derived from the years of research in Psychology) and Principal Component Analysis (PCA) would provide insight into dimension reduction techniques. Finally, this will also help identify a suitable classification technique.

5.1 Classification based on raw TIPI survey

For the first part of the classification, the raw inputs from users were used as features. Based on the multi-class nature of the target class (one of 14 general categories of academic majors), a decision tree was used for classification.

Since each user was asked for their preferred major three times, there are three answers for each participant. A target list consisting of these three major preferences was built, and the table is stacked so that the number of total rows is tripled (as each user had three answers).

Feature List	Target
gender, 10 questions asked for personality traits questionnaire	Management, Business, Health Service Law / Administration, Education Engineering, Social and Behavioural Science Life Science, Language and Literature Vocational, Arts, Communication & Media Physical Science, Philosophy & Theology

Table 5.1: Features and target class for classification with raw TIPI survey responses

5.1.1 Features and Target class

Scikit-learn [22] library was used for decision-tree classification. The features and target are as shown in the table 5.1

5.1.2 Tree depth and classification

The decision tree is also affected by the depth of the tree - hence it is benchmarked with the tree of depth 1 to 100 level. This allows for finding optimal tree depth. Figure 5.1 shows the accuracy of classification over different depth of the decision tree and using different feature sets. The red graph is using raw answers, which generally performs the best.

From the graph and list of accuracy, the decision tree performs really well **at a depth of 16 with an accuracy of 0.959**. While the accuracy may slightly increase for a high depth of the tree; we want to keep the tree as shallow as possible. It has highly diminishing returns for depth over 20. Keeping the tree shallow also ensures that the model does not suffer from overfitting, keeping the model generalizable.

5.2 Dimension Redaction - by OCEAN Indexes

For the larger dataset, using the answer of each TIPI questionnaire as the feature set to train the classifier is often time consuming and unnecessary. This can be done with a similar accuracy by the technique of dimension reductions. Most of the time in machine learning, dimension reduction is a black box, and the feature sets derived from a higher dimension to lower dimensions have no intuitive real-world meaning. We use OCEAN Index, which

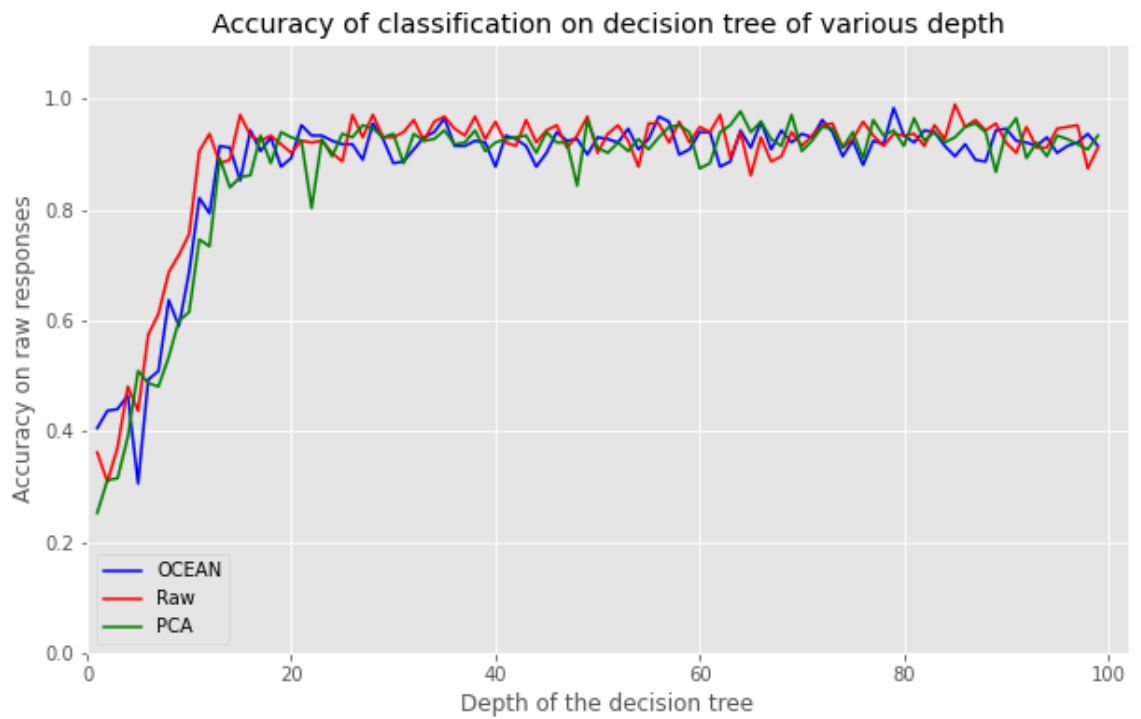


Fig. 5.1: Accuracy of the decision tree over different tree depth using different features

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
mean	5.05	5.31	3.8	4.9	4.79
std	1.25	1.33	1.5	1.3	1.48
min	1	1.5	1	1	1
25th percentile	4	4.5	2.5	4	4
50th percentile	5	5.5	4	5	5
75th percentile	6	6.5	5	6	6
max	7	7	7	7	7

Table 5.2: The aggregate statistical details of calculated OCEAN Indices - MTurk Survey

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
mean	5.13	5.8	4.47	5.33	4.70
std	0.99	1.05	1.3	1.01	1.37
min	3	2	1	2.5	1
25th percentile	4.5	5	3.5	4.5	3.88
50th percentile	5	6	4.5	5.5	5
75th percentile	6	6.5	5.5	6	6
max	7	7	7	7	7

Table 5.3: The aggregate statistical details of calculated OCEAN Indices - Utah State University Survey

reduces 10 TIPI questions to 5 interpretable attributes.

5.2.1 Personality Type calculations

For the data summarized in Chapter 3, calculations were done as described in the experiment design to get the personality type for each user.

The brief statistical properties of observed personality types, in MTurk data and USU data are shown in the table 5.2 and table 5.3 respectively:

5.2.2 Features and Target class

The five features derived from the social science researches in the form of the OCEAN index along with the gender were used as the features here. They are listed in table 5.7.

5.2.3 Tree depth and classification

The decision tree is also affected by the depth of the tree - hence it is benchmarked

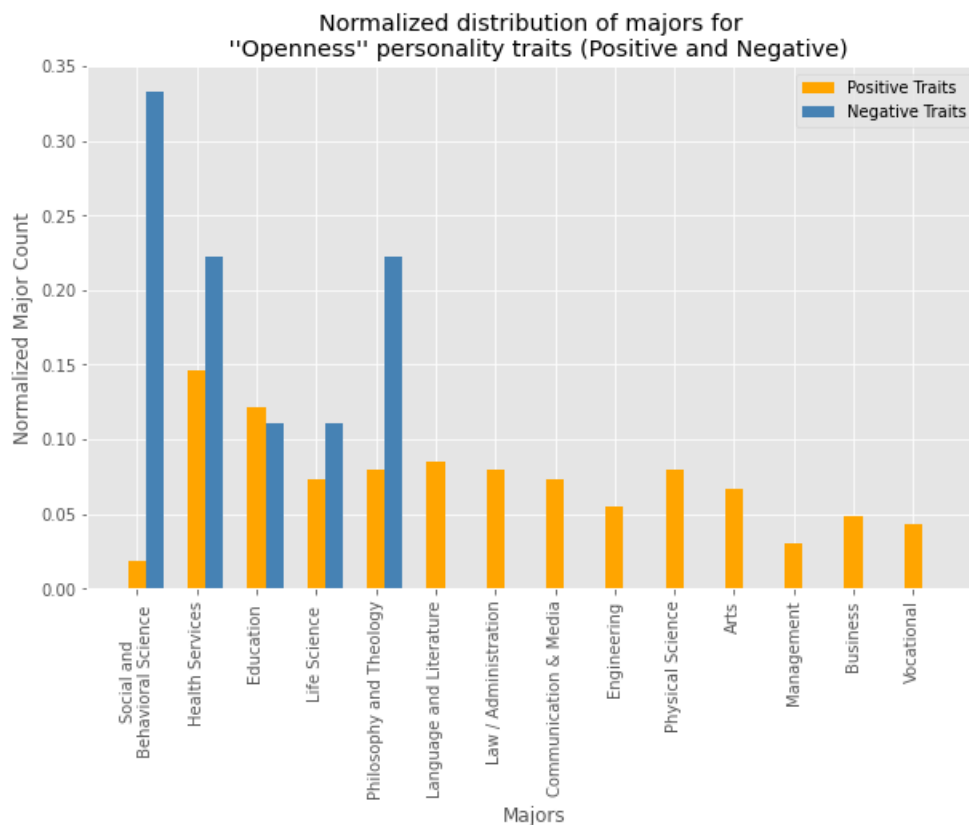


Fig. 5.2: Distribution of college majors based on Openness using results from the MTurk Survey

with the tree of depth 1 to 100 level. This allows for finding optimal tree depth. Figure 5.1 shows the accuracy of classification over different depth of the decision tree.

From the graph and list of accuracy, the decision tree performs really well **at a depth of 17 at the accuracy of 0.92 and later at dept 20 with an accuracy of 0.95** Here, we can see that the number of feature sets was cut in half; however, the accuracy is still very close to when compared with using the raw data.

5.2.4 College Majors for different personality traits

More detailed exploration was done to see the distribution of college majors for positive and negative scale of all five personality types. Figures 5.2 , 5.3, 5.4, 5.5, 5.6 show the distribution in MTurk data.

The Utah State University Survey is attached in Appendix D

Feature List	Target
gender, Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism	Management, Business, Health Service Law / Administration, Education Engineering, Social and Behavioural Science Life Science, Language and Literature Vocational, Arts, Communication & Media Physical Science, Philosophy & Theology

Table 5.4: Features and target class for classification with raw data

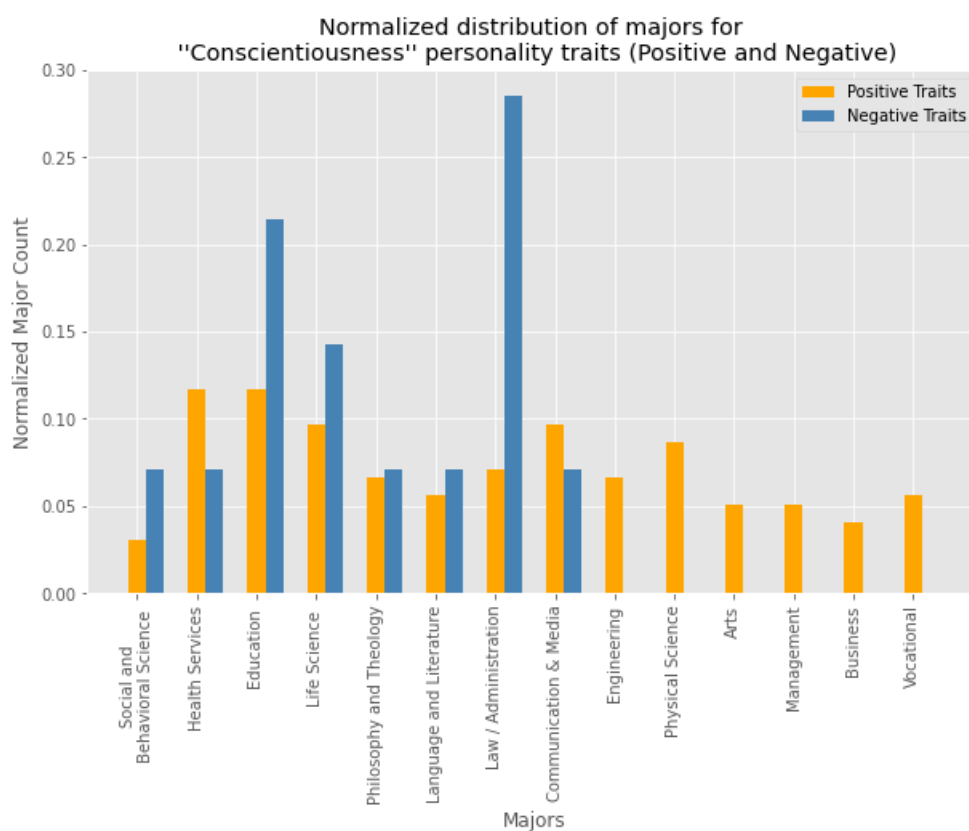


Fig. 5.3: Distribution of college majors based on Conscientiousness using results from the MTurk Survey

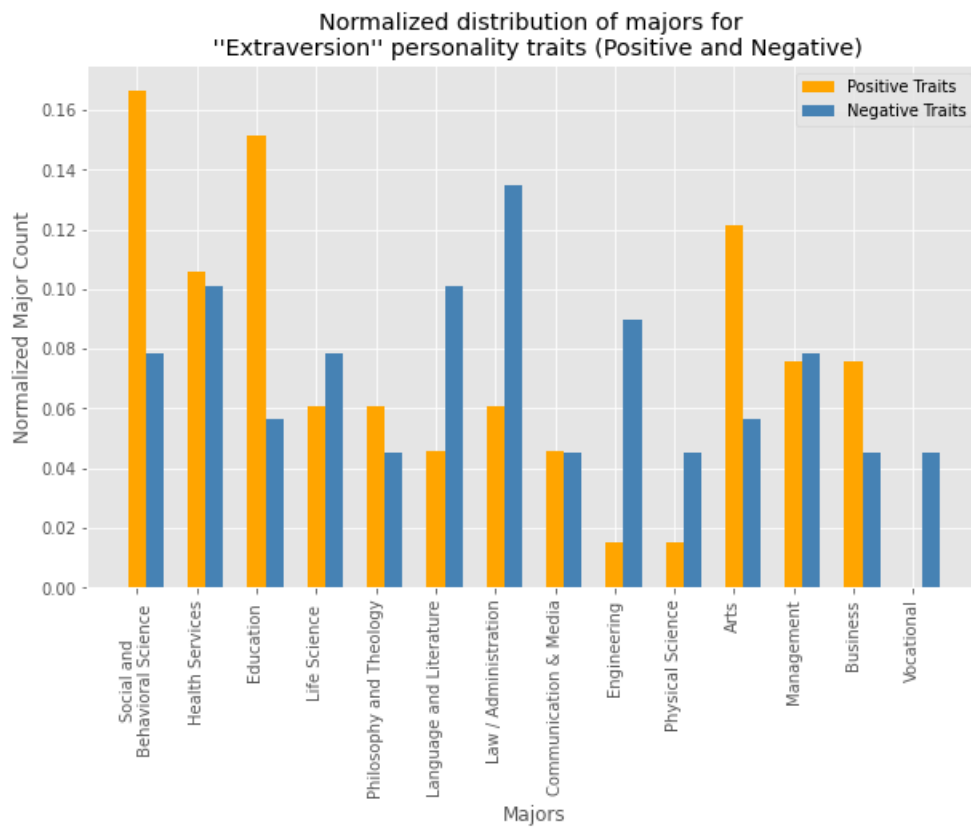


Fig. 5.4: Distribution of college majors based on Extraversion using results from the MTurk Survey

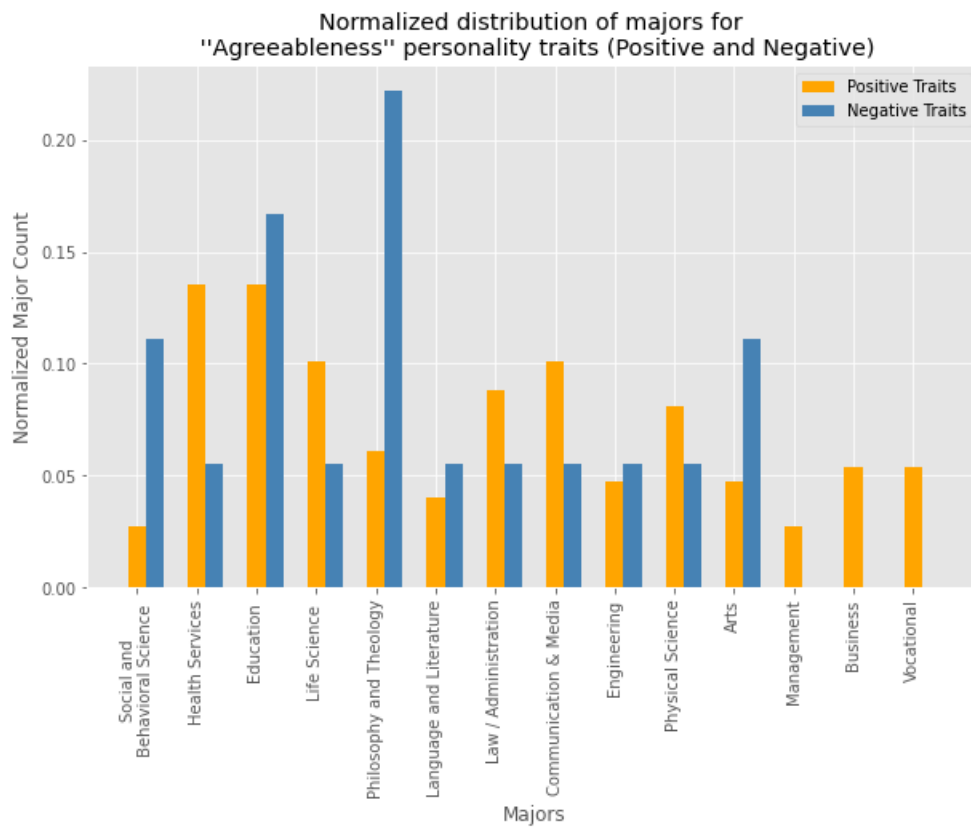


Fig. 5.5: Distribution of college majors based on Agreeableness using results from the MTurk Survey

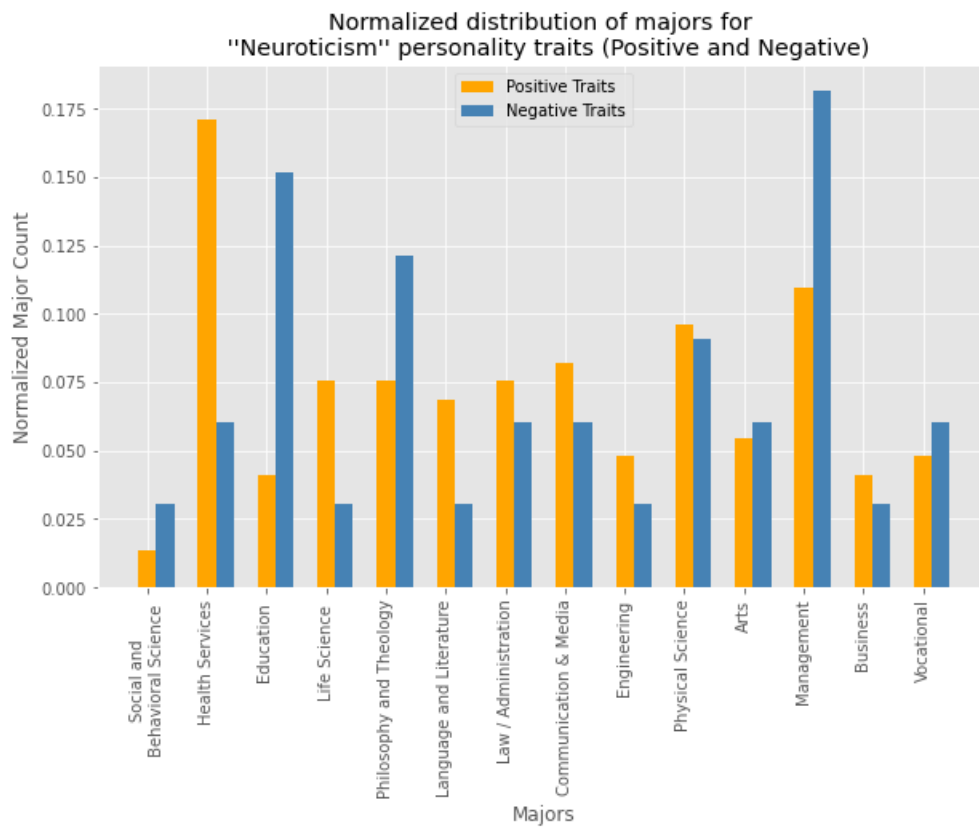


Fig. 5.6: Distribution of college majors based on Neuroticism using results from the MTurk Survey

Feature List	Target
gender,	Management, Business, Health Service
Principal Component 1,	Law / Administration, Education
Principal Component 2,	Engineering, Social and Behavioural Science
Principal Component 3,	Life Science, Language and Literature
Principal Component 4,	Vocational, Arts, Communication & Media
Principal Component 5	Physical Science, Philosophy & Theology

Table 5.5: Features and target class for classification with raw data

5.3 Dimension Reduction by Principal Component Analysis (PCA)

For the third technique, Scikit-learn's Component Analysis (PCA) was used for reducing the dimension of user responses into five components. Unlike OCEAN, these components do not have a real-world meaning. However, they represent the 10 TIPI questions.

5.3.1 Features and Target class

The data was first fit into PCA to get the first five principal components. Scikit-learn was used for decision-tree classification. The features and target are as shown in the table

5.5

```
rawFeatures = StandardScaler().fit_transform(rawFeatures)
pca = decomposition.PCA(n_components=5)
principalComponents = pca.fit_transform(rawFeatures)
principalDf = pd.DataFrame(data = principalComponents
    , columns = ['pc1', 'pc2', 'pc3', 'pc4', 'pc5'])
```

5.3.2 Tree depth and classification

The decision tree is also affected by the depth of the tree - hence it is benchmarked with tree of depth 1 to 100 level. This allows for finding optimal tree depth. Figure 5.1 shows the accuracy of classification over different depth of the decision tree.

From the graph and list of accuracy, the decision tree performs really well **at a depth of 20 at the accuracy of 0.94.**

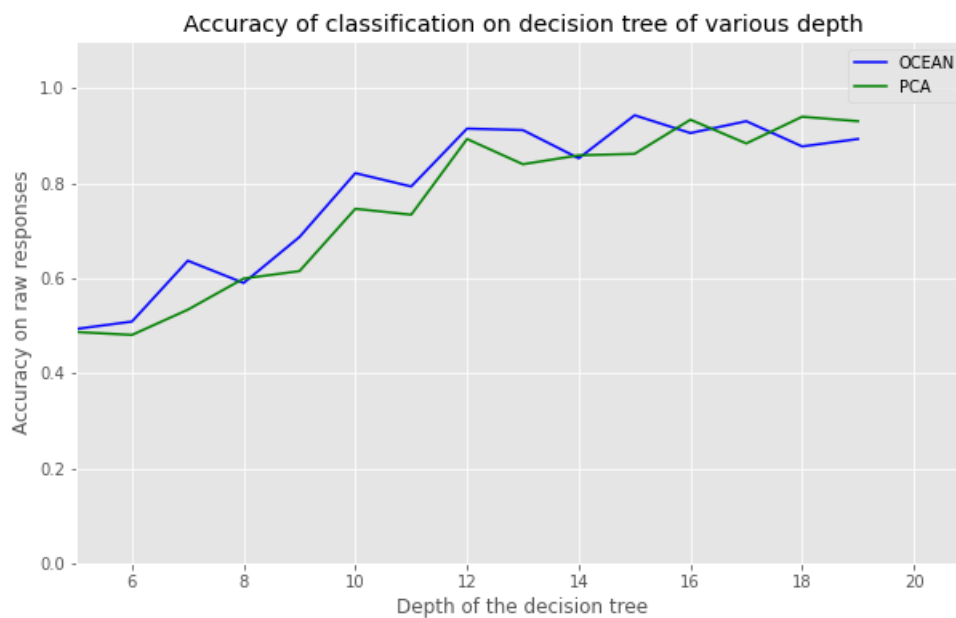


Fig. 5.7: Accuracy of OCEAN and PCA technique between depth 5-20

5.4 Accuracy among Raw, OCEAN Dimension Reduced and PCA Dimension Reduced Features

With enough tree depth, these accuracies mostly converse. There is not much of a difference between them, as evident in the figure 5.1

However, if zoomed in and looked for the tree depth between 5 to 15, the PCA lags slightly behind OCEAN dimensions, but it eventually catches up with enough depth. This can be seen in figure 5.7

5.5 Classification with Neural Network

The whole classification was repeated with the Neural Network classification technique using MLPClassifier. This model optimizes the log-loss function using LBFGS, or stochastic gradient descent [22]. The classification was done using different solver, and parameters used are shown in table 5.6 This model **achieved an accuracy of 0.66 with over 400 hidden neurons**. While that is a moderately good result for a 14-class target variable, this was nowhere near the accuracy of the decision tree. Similarly, it took a lot longer to train. In

an 8-core 4.4 GHz processor machine, training in neural networks with 500 hidden neurons took over 21 minutes, while the decision tree is unusually trained within 10 seconds.

Parameters	Value
Hidden Layers	1 to 250 (default= 100)
solver	'lbfgs', 'sgd', 'adam'
alpha	1e-5
activation	Relu

Table 5.6: Parameters for neural network MLPClassifier

The accuracy of neural Networks across the different number of hidden neurons is shown in the figure 5.8

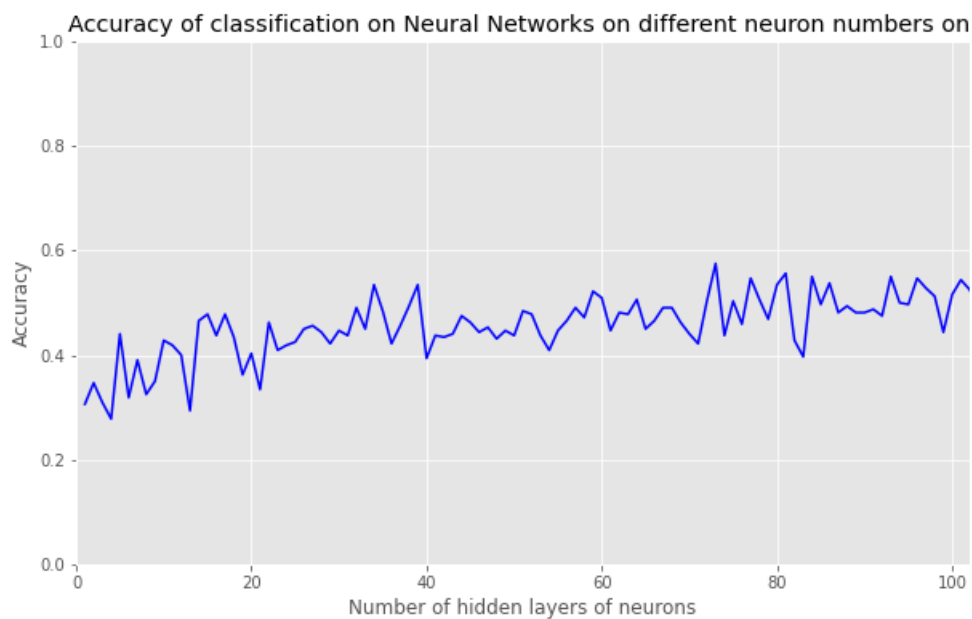


Fig. 5.8: Accuracy of the different decision tree over different tree depth

Feature List	Target
gender, Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism	Management, Business, Health Service Law / Administration, Education Engineering, Social and Behavioural Science Life Science, Language and Literature Vocational, Arts, Communication & Media Physical Science, Philosophy & Theology

Table 5.7: Features and target class for Neural Networks Classification

Traits	T-test value	P-value
Openness	0.74	0.45
Conscientiousness	4.91	1.01^{-06}
Extraversion	5.6	2.36^{-08}
Agreeableness	4.3	1.38^{-05}
Neuroticism	0.71	0.47

Table 5.8: The result of T-Test between the personality traits derived from two dataset

5.6 Classification on Utah State University Survey Data

The classifier was retrained in the training split (70 percent) of the Utah State University data before making recommendations for USU surveys and benchmarked with the 30 percent of test data. Due to sample bias, the accuracy was poor when tested without retraining. This shows that the classifier should not be generalized to any particular group or demography. It is very important to train it in the same kind of training data. Even with the slightly skewed sampling, the USU Survey data performed roughly in the same accuracy as MTurk data when the model is retrained and benchmarked. The accuracy of a decision tree with the raw survey, dimension reduced by OCEAN index and PCA technique are attached in the figure 5.9.

A pairwise T-Test was done on the OCEAN score responses of MTurk and USU data. The result of T-test is in the table 5.8.

Put all together, the USU survey data could be classified just as well as MTurk data. The accuracy of all three techniques is as shown in the figure 5.9.

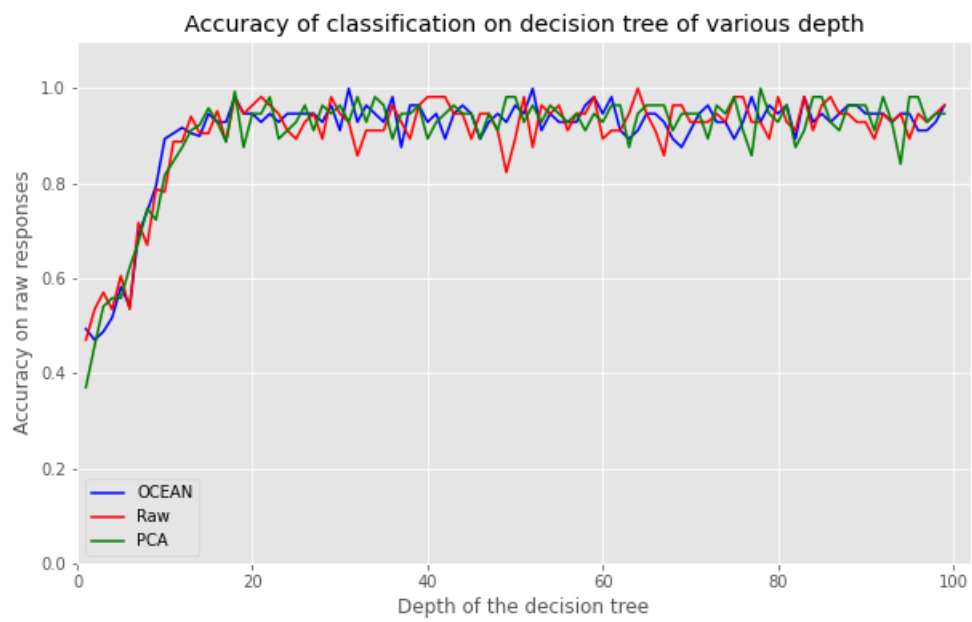


Fig. 5.9: Accuracy of the different decision tree over different tree depth

CHAPTER 6

DISCUSSION AND CONCLUSION

We began this research with three questions:

1. R1: What is the viability of the use of the personality traits for the data-driven academic recommendation?
2. R2: How do the use of expert's prior knowledge used for dimension reduction compare to the use of data science's approach?
3. R3: What is the optimal classification technique for this recommendation system?

For this study, a framework to collect the user data simulating the user onboarding process of the deployed app was created, and the survey was launched on Amazon Mechanical Turk (MTurk) platform and also through in-university recruitment. Over 800 surveys were collected, and 543 were deemed valid and analyzed.

6.1 Key Contributions

The key contribution of this research is discussed below.

6.1.1 Viability of use of the personality traits for the data-driven academic recommendation

Based on both survey, it can be inferred that people with different personality traits prefer different academic majors. Based on those results, there is a clear correlation between the personality type and major preference and can be incorporated into the recommendation system. With the framework in place, more data can be added to the system, and the confidence in the recommendation can be increased with reinforcement learning. With over 90% correct classification of major preference in both MTurk and USU data, it showed that data is resilient to some class imbalance and can be used in real-world product deployment.

So far, even though there has been some study in the relationship between academic field and personality, there has been no such study that explores the correlations across all the majors. Also, the application of the machine learning classifier to make personality traits based college recommendation is a novel application of this relationship. This could add a new factor and layer to existing college recommendation practices.

While these prediction can reinforce the existing notion on one's mind on what they should be studying, this can be looked from a different prospective and used as diagnostic and correcting tool. If we see a large number of females are recommended a certain majors because they answer personality traits question in a peculiar way, this can be used to reach out to them and organize program that can introduce them to other career paths and choices.

6.1.2 Effect of dimension reduction technique in the recommendations

In this research, two different way of dimensions reduction was used and compared. First, it was clearly demonstrated that the reduction of dimension in behavioral data could be done without the loss in classification accuracy. The ten-question survey was reduced to five feature sets, and very comparable accuracy was maintained. Of the technique itself, the social-science based technique to use Five Major Personality traits (OCEAN index) worked as well as the state-of-the-art Principal Component Analysis (PCA) technique. However, using OCEAN gives some more valuable information that is actionable. For example, if someone scores very low in 'Openness' - it can be understood that person needs some help in that area. But if only PCA is used, the components are arbitrary, and action items cannot be inferred. As per the scope of the study, the performance of PCA is being measured for this specific set of attributes and classification and has to be taken as such. This is not a study to benchmark PCA against another dimension reduction technique in general. Because of the fact that the OCEAN index and these questionnaires were derived from years of study in social science research, it performs very well in this use case.

6.1.3 Identification of usable classification technique for recommendation

With this project, a well-performing classifier was identified for the particular use case. A Decision tree classifier, with the use of OCEAN score as the dimension reduction technique performs well and also provides an insights that can be used for other use cases. This classifier achieves over 90% accuracy with relatively small training data and training time. Other classification technique like Deep Neural Network (DNN) was also used. While DNN achieved usable accuracy of 65 percent, it took more than 120 times more time to train the neural network to get that accuracy. Neural networks took over 20 minutes for 500 hidden neurons (with an accuracy of 65 percent), while decision tree classifier training took less than 10 seconds. Our goal here is not to perform an absolute benchmarking and comparison of different classifier, but to find the classifier that works best for the data we have. Decision tree has the best performance for the data size we have, but this could change if there is different data, and can be evaluated continuously.

With these three questions of the research scope answered, this can be incorporated with other planning tools to help both high-school students to align with their interests and for the guardians and guidance counselor to help facilitate their work. Moreover, when used as a part of the larger career compass suite as intended, this study makes a meaningful contribution.

6.2 Future Works

There were some limitation and caveats in this project because of scope and availability of data. One portion of data was heavily skewed to white and female demography, which can create a bias and results can not be realizable in a different context. Similarly, because the short TIPI questionnaire were used in this survey, it can have biases on self perceived traits and self reporting. A future work which a broader data set and a larger questionnaire can help reduce that biases.

In this research, the survey respondents are either young high school graduates and early college students. While surveying them gives a good insight in interest level on certain college majors - It does not prove the success in their career field. A natural future work

in the area could be a study that covers the professional in a different field and different majors. While this would be an enormous task to gather the representative data, this would give a much clearer picture of career success in a different field for people of different personality traits.

REFERENCES

- [1] E. R. Eide, M. J. Hilmer, and M. H. Showalter, "Is it where you go or what you study? The relative influence of college selectivity and college major on earnings," *Contemporary Economic Policy*, vol. 34, no. 1, pp. 37–46, jan 2016.
- [2] J. C. Weidman, "Academic Disciplines: Holland's Theory and the Study of College Students and Faculty (review)," *The Journal of Higher Education*, vol. 76, no. 2, pp. 232–234, 2005.
- [3] A. S. C. Association *et al.*, "State-by-state student to counselor ratio report: 10 years trends," 2015.
- [4] B. Hussar, J. Zhang, S. Hein, K. Wang, A. Roberts, J. Cui, M. Smith, F. B. Mann, A. Barmer, and R. Dilig, "The condition of education 2020. nces 2020-144." *National Center for Education Statistics*, 2020.
- [5] N. R. Center, "National six-year and eight-year college completion rates."
- [6] E. Blom, M. Rainer, and M. Chingos, "Comparing colleges' graduation rates," 2020.
- [7] L. J. Schneider and T. D. Overton, "Holland personality types and academic achievement." *Journal of Counseling Psychology*, vol. 30, no. 2, p. 287, 1983.
- [8] A. Allred, M. Granger, and T. Hogstrom, "The Relationship Between Academic Major, Personality Type, and Stress In College Students," *Eukaryon*, vol. 9, no. March, 2013.
- [9] D. E. Giacominio and M. D. Akers, "An examination of the differences between personal values and value types of female and male accounting and nonaccounting majors," *Issues in Accounting Education*, vol. 13, no. 3, p. 565, 1998.
- [10] C. D. Pringle, P. B. Dubose, and M. D. Yankey, "Personality Characteristics and Choice of Academic Major: Are Traditional Stereotypes Obsolete?" *College Student Journal*, vol. 44, no. 1, p. 131, 2010. [Online]. Available: <http://www.library.umaine.edu/auth/EZProxy/test/authej.asp?url=http://search.ebscohost.com/login.aspx?direct=true{%&}db=f5h{%&}AN=48646435{%&}site=ehost-live>
- [11] E. C. Tupes and R. E. Christal, "Recurrent personality factors based on trait ratings," *Journal of personality*, vol. 60, no. 2, pp. 225–251, 1992.
- [12] J. M. Digman, "Personality structure: Emergence of the five-factor model," *Annual review of psychology*, vol. 41, no. 1, pp. 417–440, 1990.
- [13] L. R. Goldberg *et al.*, "A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models," *Personality psychology in Europe*, vol. 7, no. 1, pp. 7–28, 1999.

- [14] R. J. Cebula and J. Lopes, "Determinants of Student Choice of Undergraduate Major Field," *American Educational Research Journal*, vol. 19, no. 2, pp. 303–312, 1982.
- [15] J. L. Zheng, K. P. Saunders, M. C. Shelley, and D. F. Whalen, "Predictors of academic success for freshmen residence hall students," *Journal of College Student Development*, vol. 43, no. 2, pp. 267–283, 2002.
- [16] L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. G. Gough, "The international personality item pool and the future of public-domain personality measures," *Journal of Research in personality*, vol. 40, no. 1, pp. 84–96, 2006.
- [17] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr, "A very brief measure of the big-five personality domains," *Journal of Research in personality*, vol. 37, no. 6, pp. 504–528, 2003.
- [18] C. S. Bruck and T. D. Allen, "The relationship between big five personality traits, negative affectivity, type a behavior, and work–family conflict," *Journal of vocational behavior*, vol. 63, no. 3, pp. 457–472, 2003.
- [19] A. Joshi, S. Kale, S. Chandel, and D. K. Pal, "Likert scale: Explored and explained," *Current Journal of Applied Science and Technology*, pp. 396–403, 2015.
- [20] D. of Education, "Collegescorecard.ed.gov. 2020. college scorecard data," 2020. [Online]. Available: <https://collegescorecard.ed.gov/data/>
- [21] U. S. C. Bureau, "Quickfacts," 2016.
- [22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

APPENDICES

APPENDIX A
Survey Questionnaire

Default Question Block

Please enter your age:

Please choose your sex:

- Male
- Female
- Non-binary

Please choose your race/ethnicity:

- White
- Black or African American
- American Indian or Alaska Native
- Asian
- Native Hawaiian or Other Pacific Islander
- Others

Block 1

I see myself as extraverted and enthusiastic

- Strongly agree
- Agree
- Somewhat agree
- Neither agree nor disagree

- Somewhat disagree
- Disagree
- Strongly disagree

I see myself as critical and quarrelsome

- Strongly agree
- Agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Disagree
- Strongly disagree

I see myself as dependable and self-disciplined

- Strongly agree
- Agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Disagree
- Strongly disagree

I see myself as anxious and easily upset

- Strongly agree
- Agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Disagree

Strongly disagree

I see myself as complex and open to new experiences

Strongly agree

Agree

Somewhat agree

Neither agree nor disagree

Somewhat disagree

Disagree

Strongly disagree

I see myself as reserved and quiet

Strongly agree

Agree

Somewhat agree

Neither agree nor disagree

Somewhat disagree

Disagree

Strongly disagree

I see myself as sympathetic and warm

Strongly agree

Agree

Somewhat agree

Neither agree nor disagree

Somewhat disagree

Disagree

Strongly disagree

I see myself as disorganized and careless

- Strongly agree
- Agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Disagree
- Strongly disagree

I see myself as calm and emotionally stable

- Strongly agree
- Agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Disagree
- Strongly disagree

I see myself as conventional and uncreative

- Strongly agree
- Agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Disagree
- Strongly disagree

Block 2

Of the college majors shown below, select the one that interests you the most:

- Language and Literature (Literature, English, Foreign Language, etc.)
- Life Science (Biology, Ecology, Neuroscience, etc)
- Education (Elementary Education, Special Ed and Teaching, Curriculum Development, etc)
- Health Services (Medical Doctor, Nursing, Dental, etc)
- Management (Business Administration, Finance, Management Science, etc)

Of the college majors shown below, select the one that interests you the most:

- Social and Behavioral Science (Psychology, Human Development, History, etc)
- Philosophy and Theology (Philosophy, Religious Education, Ethics, etc)
- Management (Business Administration, Finance, Management Science, etc)
- Physical Science (Physics, Mathematics, Chemistry, etc)
- Language and Literature (Literature, English, Foreign Language, etc.)

Of the college majors shown below, select the one that interests you the most:

- Arts (Music, Fine Art, Theater, etc)
- Life Science (Biology, Ecology, Neuroscience, etc)
- Management (Business Administration, Finance, Management Science, etc)
- Communication & Media (Journalism, Public Relations, Social Media Marketing, etc)
- Law / Administration (Criminal Justice, Law, Political Science, etc)

What is answer to $2 + 3$?

- 4
- 55
- 5
- 7

Block 3

Here is your Survey Completion code: \${e://Field/Random%20ID}
Enter this code to the Amazon Mechanical Turk to complete the survey.



Please make sure you hit the 'NEXT' button below to record your survey responses.

Powered by Qualtrics

APPENDIX B
IRB Exemption Approval



Exemption #2
Certificate of Exemption

From: Melanie Domenech Rodriguez, IRB Chair 
Nicole Vouvalis, IRB Director 

To: **Travis Dorsch**

Date: **October 15, 2020**

Protocol #: **11463**

Title: **Linkages between personality type and college major**

The Institutional Review Board has determined that the above-referenced study is exempt from review under federal guidelines 45 CFR Part 46.104(d) category #2:

Research that only includes interactions involving educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior (including visual or auditory recording) if at least one of the following criteria is met: (i) The information obtained is recorded in such a manner that the identity of the human subjects cannot readily be ascertained, directly or through identifiers linked to the subject; (ii) Any disclosure of the responses outside the research would not reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, educational advancement, or reputation, or (iii) the information obtained is recorded by the investigator in such a manner that the identity of the human subjects can readily be ascertained, directly or through identifiers linked to the subjects, and the IRB conducts a limited IRB review to make required determinations.

This study is subject to ongoing COVID-19 related restrictions. As of March 15, 2020, the IRB has temporarily paused all in person research activities, including but not limited to recruitment, informed consent, data collection and data analysis that involves personal interaction (such as member checking and meaning-making). If research cannot be paused, please file an amendment to your protocol modifying procedures that are conducted in person. The IRB will notify you when in person research activities are once again permitted.

This exemption is valid for five years from the date of this correspondence, after which the study will be closed. If the research will extend beyond five years, it is your responsibility as the Principal Investigator to notify the IRB **before** the study's expiration date and submit a new application to continue the research. Research activities that continue beyond the expiration date without new certification of exempt status will be in violation of those federal guidelines which permit the exempt status.

If this project involves Non-USU personnel, they may not begin work on it (regardless of the approval status at USU) until a Reliance Agreement, External Research Agreement, or separate protocol review has been completed with the appropriate external entity. Many schools will not engage in a Reliance Agreement for Exempt protocols, so the research team must determine what the appropriate approval mechanism is for their Non-USU colleagues. As part of the IRB's quality assurance procedures, this research may be randomly selected for audit during the five-year period of exemption. If so, you will receive a request for completion of an Audit Report form during the month of the anniversary date of this certification.

In all cases, it is your responsibility to notify the IRB **prior** to making any changes to the study by submitting an Amendment request. This will document whether or not the study still meets the requirements for exempt status under federal regulations.

Upon receipt of this memo, you may begin your research. If you have questions, please call the IRB office at (435) 797-1821 or email to irb@usu.edu.

The IRB wishes you success with your research.

APPENDIX C
IRB Letter of Information

Linkages between personality type and college major

You are invited to participate in a research study by Dr. Travis Dorsch, an associate professor in the department of Human Development and Family Studies, Dr. John Edwards, an assistant professor, and Aashish Ghimire, a student in the Department of Computer Science at Utah State University.

The purpose of this research is to study the relationship between personality traits and preference of academic major. Specifically, we are interested in learning about what college majors people with different personality types are interested in. You are being asked to participate in this research because you opted to complete this short survey and help the research.

Your participation in this study is voluntary and you may withdraw your participation at any time for any reason.

If you take part in this study, you will be asked to complete a 16 questions survey describing yourself and your preference in college majors. This is expected to take about 3-5 minutes.

The possible risks of participating in this study include possible loss of confidentiality. We cannot guarantee that you will directly benefit from this study, but it has been designed to learn more about relation between personality traits and college majors.

We will make every effort to ensure that the information you provide remains confidential. We will not reveal your identity in any publications, presentations, or reports resulting from this research study. We will not collect any personal identifiable information such as your name, email address or IP address.

We will collect your information through an anonymous Qualtrics survey. Online activities always carry a risk of a data breach, but we will use systems and processes that minimize breach opportunities. This survey will be securely stored in encrypted USU BOX storage and hardware.

For your participation in this research study you will receive one USD or equivalent amazon gift certificate. There will be a very simple attention check verification question with an arithmetic addition, and that question needs to be answered correctly to get compensated.

You can decline to participate in any part of this study for any reason and can end your participation at any time.

If you have any questions about this study, you can contact John Edwards at john.edwards@usu.edu, Travis Dorsch at travis.dorsch@usu.edu or Aashish Ghimire at aashish.ghimire@usu.edu . Thank you again for your time and consideration. If you have any concerns about this study, please contact Utah State University's Human Research Protection Office at (435) 797-0567 or irb@usu.edu.

By continuing to the survey you agree that you are 18 years of age or older, and wish to participate. You agree that you understand the risks and benefits of participation, and that you know what you are being asked to do. You also agree that if you have contacted the research team with any questions about your participation, and are clear on how to stop your participation in this study if you choose to do so. Please be sure to retain a copy of this form for your records.

APPENDIX D

Distribution of Personality type questions - Amazon Mechanical Turk Survey

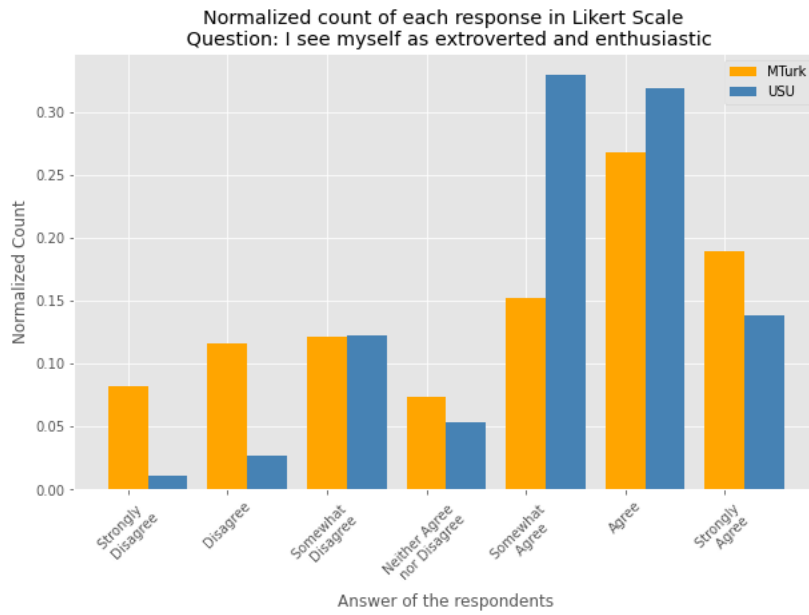


Fig. D.1: Answer distribution for the question Q1

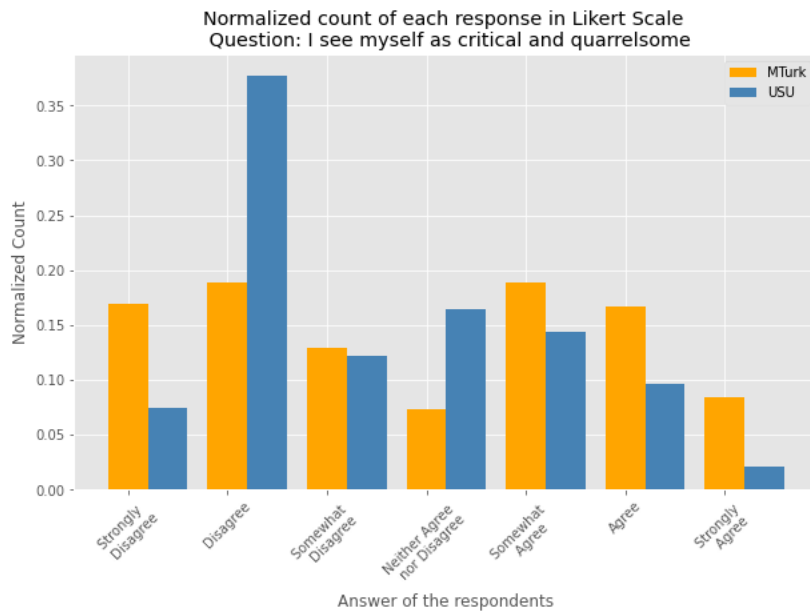


Fig. D.2: Answer distribution for the question Q2

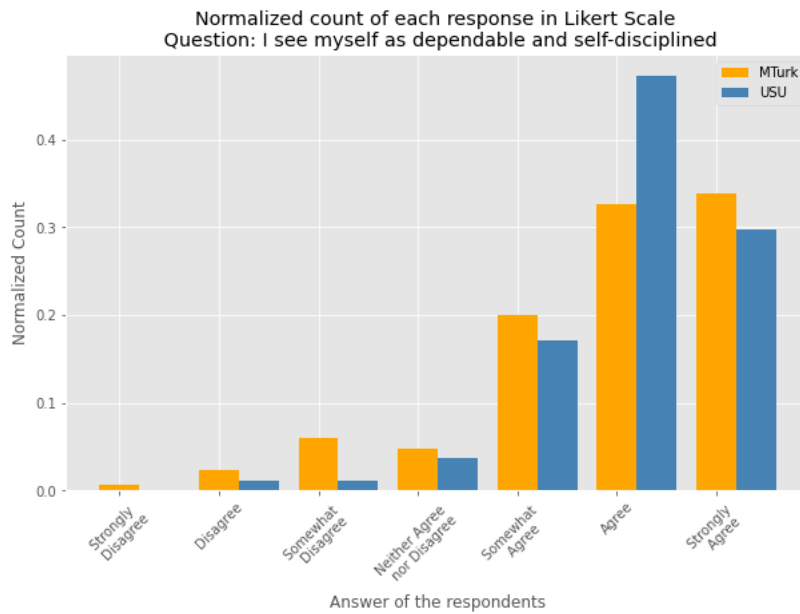


Fig. D.3: Answer distribution for the question Q3

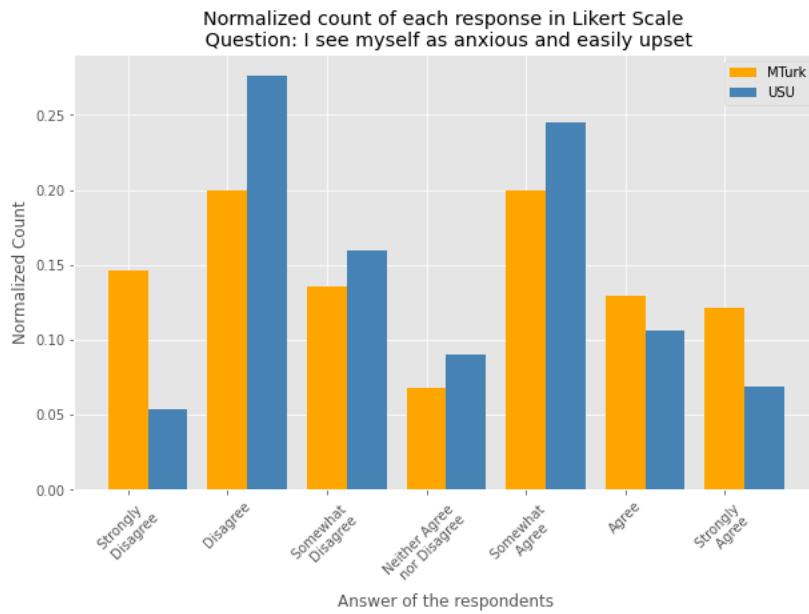


Fig. D.4: Answer distribution for the question Q4

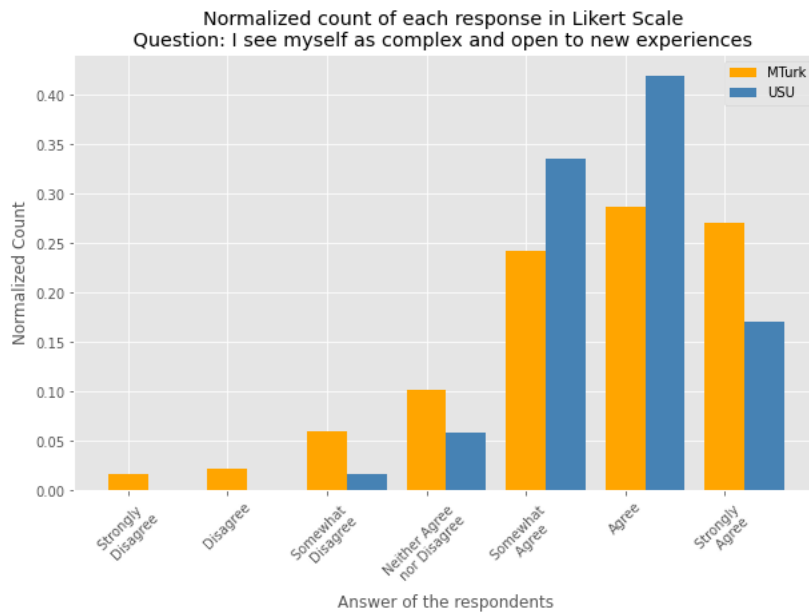


Fig. D.5: Answer distribution for the question Q5

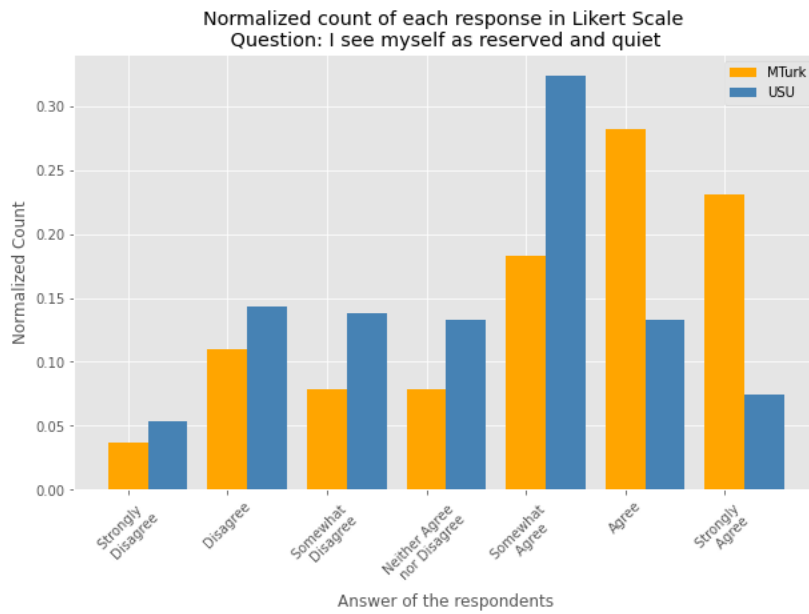


Fig. D.6: Answer distribution for the question Q6

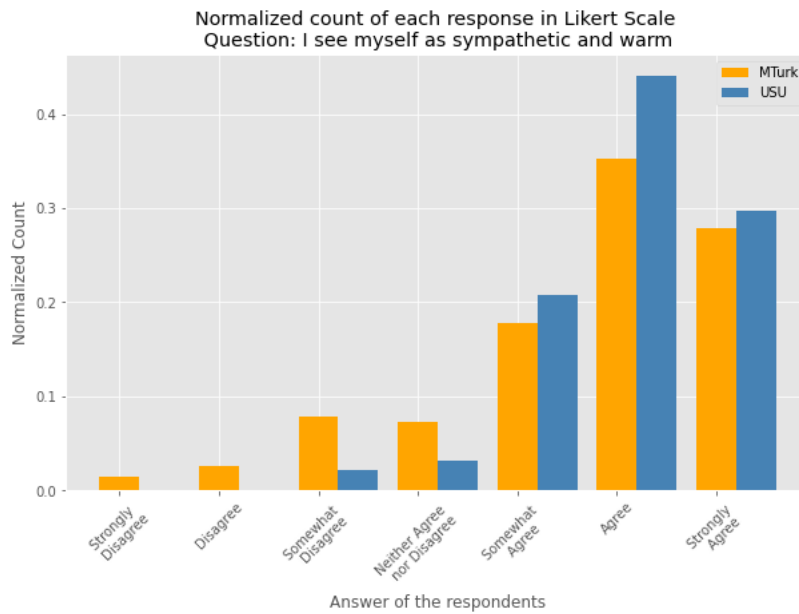


Fig. D.7: Answer distribution for the question Q7

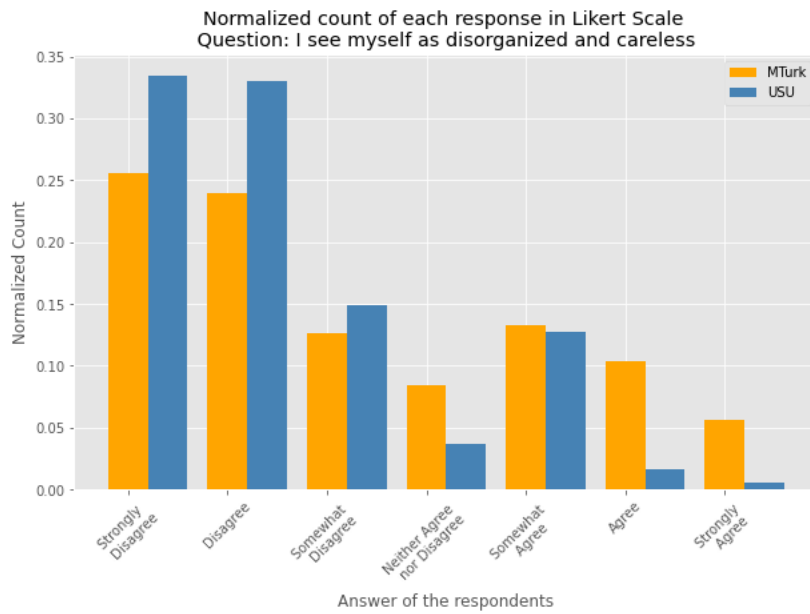


Fig. D.8: Answer distribution for the question Q8

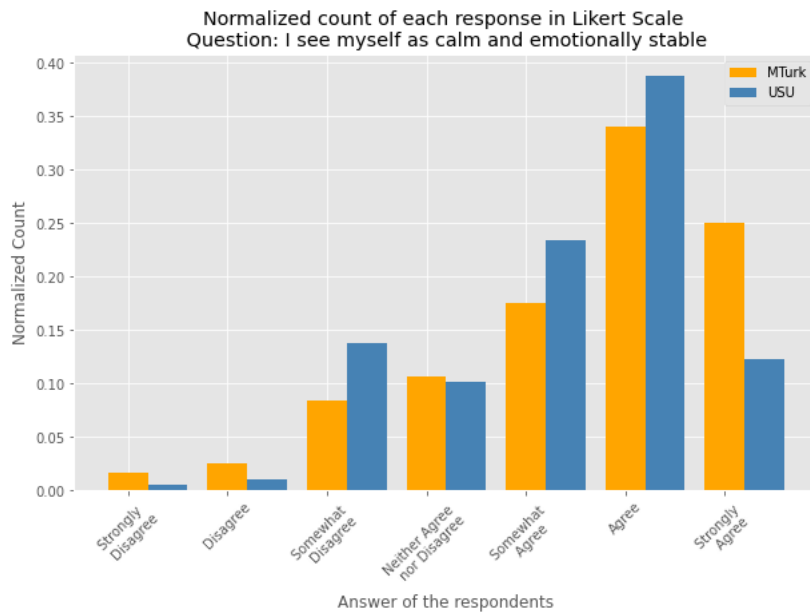


Fig. D.9: Answer distribution for the question Q9

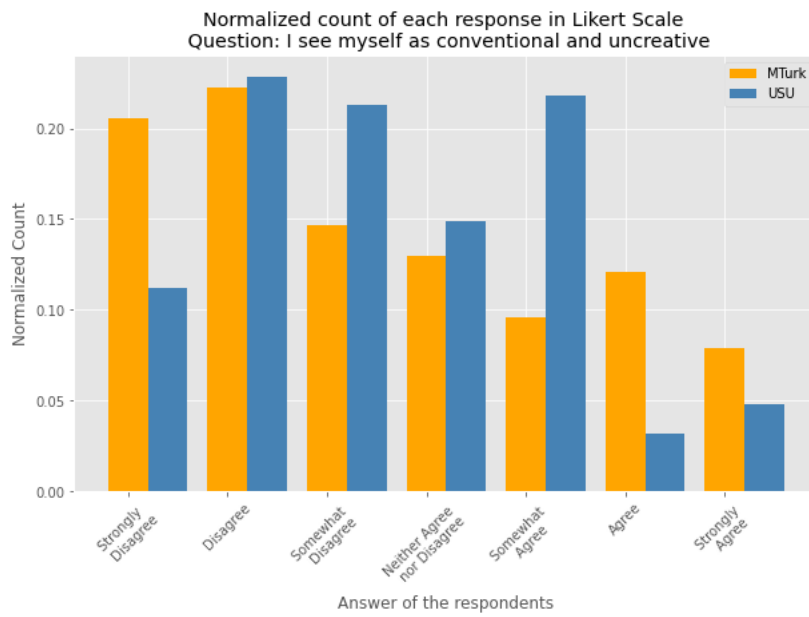


Fig. D.10: Answer distribution for the question Q10

APPENDIX E

Distribution of Major across Personality type questions - Utah State University Survey

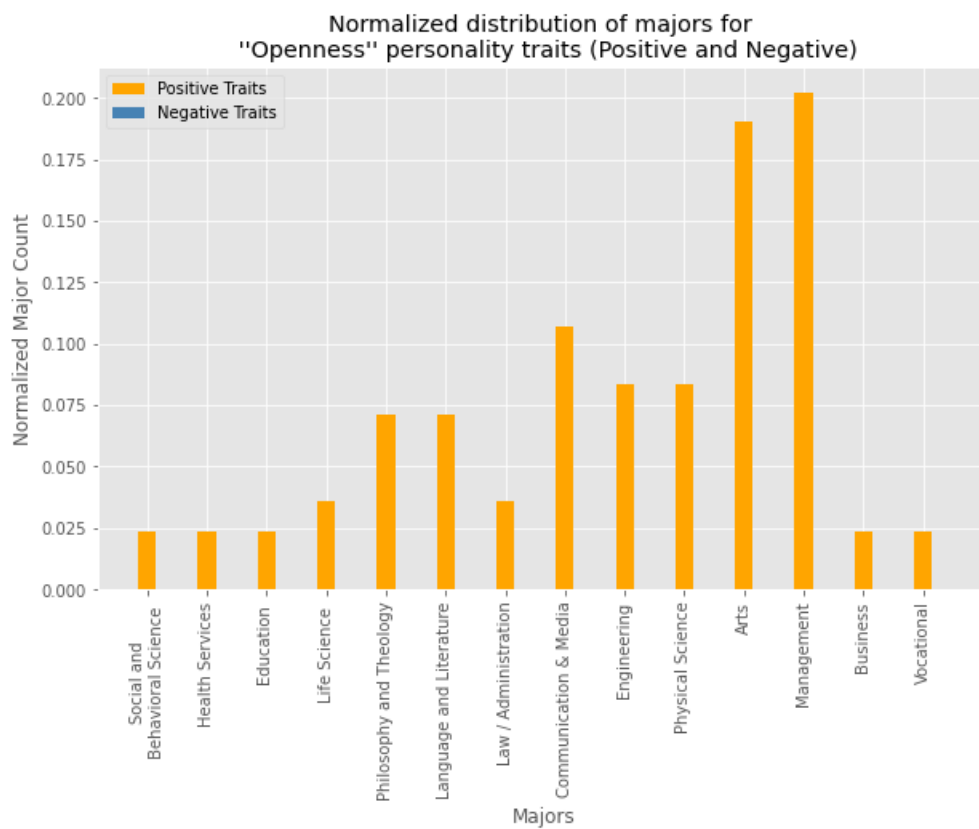


Fig. E.1: Distribution of college majors based on Openness

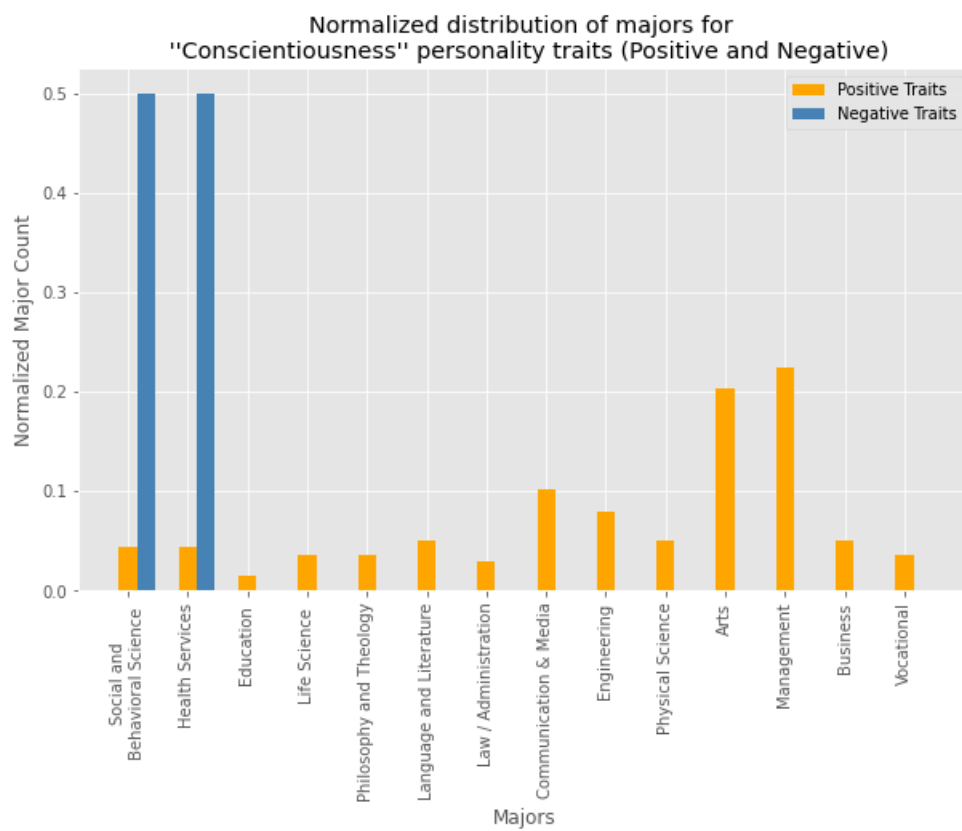


Fig. E.2: Distribution of college majors based on Conscientiousness

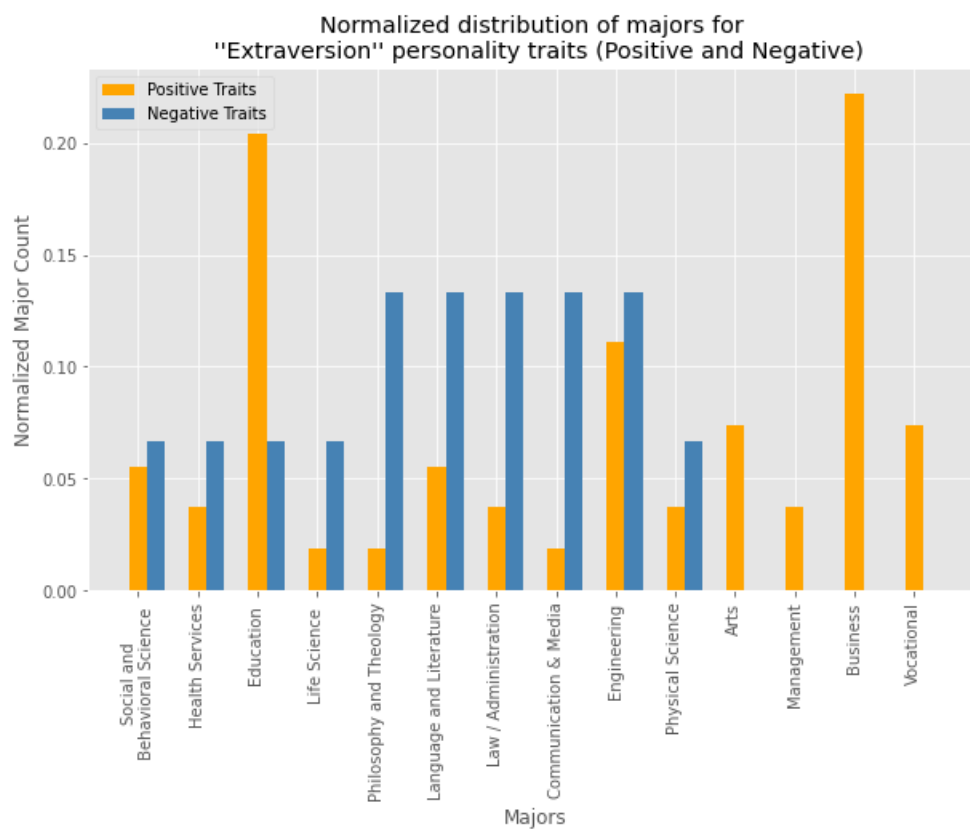


Fig. E.3: Distribution of college majors based on Extraversion

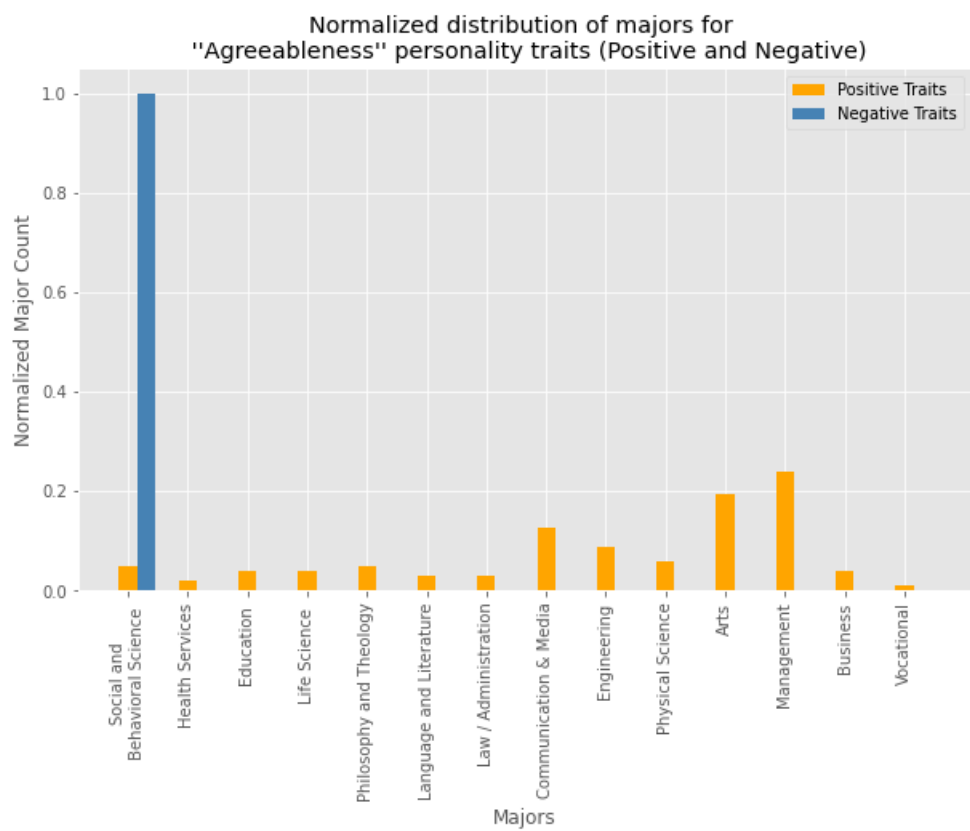


Fig. E.4: Distribution of college majors based on Agreeableness

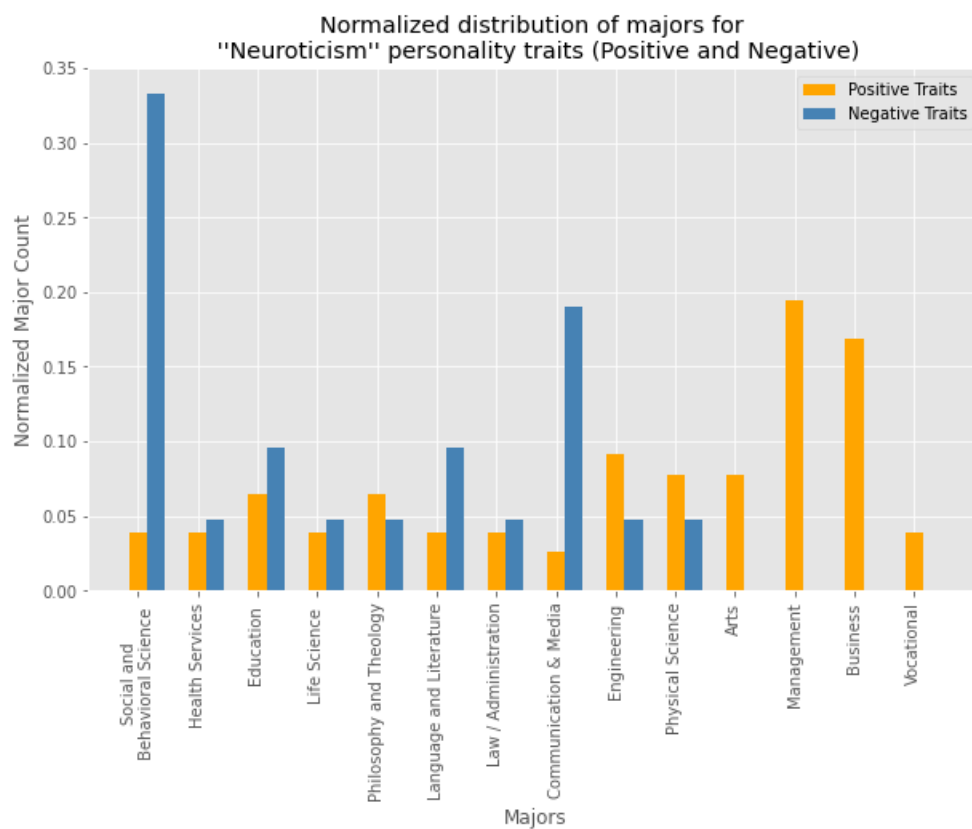


Fig. E.5: Distribution of college majors based on Neuroticism