Data-Driven Recommendation of Academic Options Based on Personality Traits

Aashish Ghimire
Department of Computer Science
Utah State University
Logan, Utah
aashish.ghimire@usu.edu

Travis Dorsch
Department of Human & Family Studies
Utah State University
Logan, Utah
travis.dorsch@usu.edu

John Edwards
Department of Computer Science
Utah State University
Logan, Utah
john.edwards@usu.edu

***Abstract*—The choice of academic major and academic institution has a large effect on a person's career. About 40% of students either transfer to a different major or different college or drop out of college within six years. 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. In light of this gap in understanding, we surveyed over 500 people between 18 and 25 years old to capture 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 a decision tree model, accurate classification can be done, with over 90% accuracy. Furthermore, we explored the two methods of dimension reduction - one using Principal Component Analysis (PCA) and another relying on Social Science research on the Big-Five personality Traits (also known as OCEAN indices) to simplify the problem further. With these techniques, the dimension was reduced by half without decreasing the accuracy of our classifier. We compared other popular machine learning methods and demonstrated that a decision tree is best for such an application. With this research, a readily deployable recommendation system was created that can help students find their most enjoyable academic path and aid guidance counselor and parents with their recommendations.**

***Index Terms*—classification, principal component analysis(pca), personality types, academic preference**

I. Introduction

The choices of academic major and institution are among the most fundamental steps in a person's career. A student typically makes that decision during their high school years 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 [1]. This 6- year limit does not include the breaks taken for military service, religious service, etc. For the cohort starting in 2011, the eight-year graduation rate was only 61.8 percent [2]. Traditionally, schools provide a guidance counselor to help students make 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 a data-driven approach. 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 there is little beyond that, according to market research done by Graphium, Inc. [4]. There are some commercial services available directly to 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.

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 [5] classified academic major with six personality types. 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 has been widely used across different papers in psychology research for decades [6]. Allred et al. studied the validity of the stereotypes surrounding the choice of academic majors and stress level in different academic disciplines [7]. Most of these findings are from qualitative studies that do not use a data driven approach to validate their conclusions. In building on these findings, our study was designed to answer three research questions:

- R1: How effective is the use of personality traits in predicting a preferred college major?
- R2: How do expert-derived personality traits compare to a data-driven dimension reduction technique?
- R3: What are the unique personality traits found in students preferring different majors?

Identify applicable funding agency here. If none, delete this.

II. Methods

Since there is no publicly available dataset suitable for classifying a student's preference of major based on their personality, we, as a major part of this study, collected such a dataset in the form of a survey. This survey asked ten probing personality questions, as listed below. It also asked three questions related to college major preference, along with demographics questions. Utilizing survey responses, the scores for five personality traits derived from the ten personality questions were used to calculate 5-dimensional personality features using the OCEAN model. For comparison, we also built classification models using 5-dimensional personality featured from the ten personality questions using Principal Component Analysis (PCA). In the present study, we compare the performance of the expert-derived OCEAN personality traits with PCA-derived traits.

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

In the user survey, the following ten questions [8] were asked to gauge the personality type of the user. Each answer is on a seven-level Likert scale:

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

B. Big Five Personality Traits (OCEAN) for dimension reduction

For classifying the user and better understanding their personality type, we use the Big Five Personality Trait index [9]. The five traits are openness, conscientiousness, extraversion, agreeableness, and neuroticism. This is often referred as the OCEAN index. In our study, each trait was classified as 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, neutral, or positive in each of the five personality traits gives interpretable categories that can be used to provide personalized guidance to the students. 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.

C. Calculations of the mean OCEAN scores

A seven-level Likert Scale [10] 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 [8]. 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., where TQ1, TQ2 ... TQ10 refer to answers to the 10 TIPI personality questions. Those formulae are shown in 1, 2...5.

$$Openness = \frac{\text{TP5 + (8 - TP10)}}{2} \tag{1}$$

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

$$Extraversion = \frac{\text{TP1 + (8 - TP6)}}{2} \tag{3}$$

$$Agreeableness = \frac{\text{(8 - TP2) + TP7}}{2} \tag{4}$$

$$Neuroticism = \frac{\text{(8 - TP4) + TP9}}{2} \tag{5}$$

D. Academic Major Preference

The list of all college majors listed by the US Department of education was obtained from their public database [11]. This yielded 397 unique majors. To narrow the list, only majors offered at more than 100 schools were picked, which left 261 majors. From there, we divided the majors into 14 general categories based on how these majors are commonly classified in their school's organization structure. These are shown in Table I. After asking three demographics and ten personality questions, survey takers were provided with an attention check question such as "what is 2 + 2?". After passing 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. An example of a question regarding the academic major in the survey is shown below:

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)

E. Survey Platform and Data Collection

The survey was designed using Qualtrics's survey platform. Survey responses were collected from two different sources: Utah State University and Amazon Mechanical Turk (MTurk) workers. For the sake of comparison, these two groups of data were analyzed and used separately. Our Institutional Review Board (IRB) approved the surveys for both Utah State University students and MTurk workers with the questionnaire varied slightly to fit the logistics of validating and paying

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 I
TABLE SHOWING THE NUMBER OF MAJORS IN EACH GENERAL MAJOR CATEGORY

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

TABLE II
GENDER DISTRIBUTION OF PARTICIPANT IN MTURK AND USU DATA.

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/Alaska Native	3	0.8%	3	1.59%
Asian	41	11.54%	15	7.97%
Native Hawaiian/Pacific Islander	1	0.28%	1	0.53%
Others	11	3.09%	3	1.59%

TABLE III
RACE DISTRIBUTION OF PARTICIPANT IN MTURK AND USU DATA.

MTurk's respondents. All participants were compensated with one dollar in the form of Amazon gift card.

1) Survey of students at Utah State University: Our survey was completed by students at Utah State University. A Quick Response (QR) Code linked to the survey was emailed to the cadets of the US Army Reserves Officer Training Corps (ROTC) program at the University. The QR code was also sent by a Teaching Assistant to students in a General Psychology class at USU. The age of the respondents was restricted to 25 and under.

2) Survey using Amazon Mechanical Turk (MTurk) platform: The MTurk platform was used to gather additional survey results. MTurk is a commercial survey and crowdsourcing platform where users are compensated for completing tasks. Using MTurk's filter feature, we restricted the respondents to only 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 spam responses, 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 work. For example, if a user tends to complete surveys in a very short time, or they use other programmatic tools to complete it, they get rejected more often and have a 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.

III. SURVEY RESULTS

A. Survey Participation: Amazon Mechanical Turk

There were a total of 728 responses from Amazon Mechanical Turk, of which 420 who fully completed the survey were aged between 18 to 25. We next looked into the completion time and removed any surveys that took less than 60 seconds. 355 responses remained after this filter. Among these 355 responses, the mean completion time was 142 seconds, and the median was 106 seconds.

1) Gender distribution: In the MTurk survey, as shown in Table II, survey participant are fairly even in gender distribution as compared to USU data.

2) Race distribution: The race distribution of survey is as listed in Table III. The MTurk distribution is not far off from the United State's national race distribution according to the United States Census Bureau [12].

B. 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.

1) Sampling Bias and separate processing of data:: As per most research done among two different communities, this survey is vulnerable to a sampling bias. The US population is about 13.4% Black or African American (Bureau, 2016), but the proportion in Utah State University's survey is under 1this is closer to Utah State's population (1.5%). The 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. MTurk's data is well balanced in terms of gender, 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). USU participants were guaranteed to be students, but the same could not be said for the MTurk survey.

IV. CLASSIFICATION RESULTS

Our first research question (R1) is: How effective is the use of personality traits in predicting a preferred college major? To answer this question, we built a machine learning model (specifically, a decision tree) to predict a students' preferred

academic major based on their personality type. We built models using different feature vectors. The first is raw 10-question personality type. The second is 5 dimensional OCEAN index and the third is the 5-dimensional index derived from Principal Component Analysis (PCA). Using these different feature sets allowed us to objectively judge the quality of the widely accepted OCEAN index and answer our second research question R2: How do expert-derived personality traits compare to a data-driven dimension reduction technique? Furthermore, using the decision tree classification model with its high degree of interpretability helps us answer our third research question R3: What are the unique personality traits found in students preferring different majors?

A. *Classification based on raw TIPI survey*

For the first part of the classification, the 10-dimensional raw inputs from users were used as features. Decision trees natively support multi-class prediction, which is convenient because our target variable, preferred major, could be one of 14 general categories of academic majors. Since each user was asked for their preferred major three times, there are three answers for each participant. The Scikit-learn [13] library was used for decision-tree classification. The features and target are as shown in IV.

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 IV
FEATURES AND TARGET CLASS FOR CLASSIFICATION WITH RAW TIPI SURVEY RESPONSES

A sample result of normalized distribution of college majors preference in the MTurk survey based on the positive and negative responses on conscientiousness and neuroticism is shown in figure 1 and 2 respectively. Neutral responses and other personality types are not shown.

1) Tree depth and classification: The decision tree is affected by the depth of the tree – so we benchmarked with trees of depth ranging from 1 to 100 levels. This allows for finding the optimal tree depth. Figure 8 shows the accuracy of classification over different depths of decision tree and using different feature sets. The red graph is using raw answers, which generally performs the best. From Figure 3 we can see that the decision tree performs well at a depth of 16 with an accuracy of 0.959. While the accuracy may slightly increase for a higher depth of tree, we want to keep the tree as shallow as possible to ensure that the model does not suffer from overfitting, thus keeping the model generalizable.

B. *Dimension Reduction - by OCEAN Indexes*

For larger datasets, 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 similar

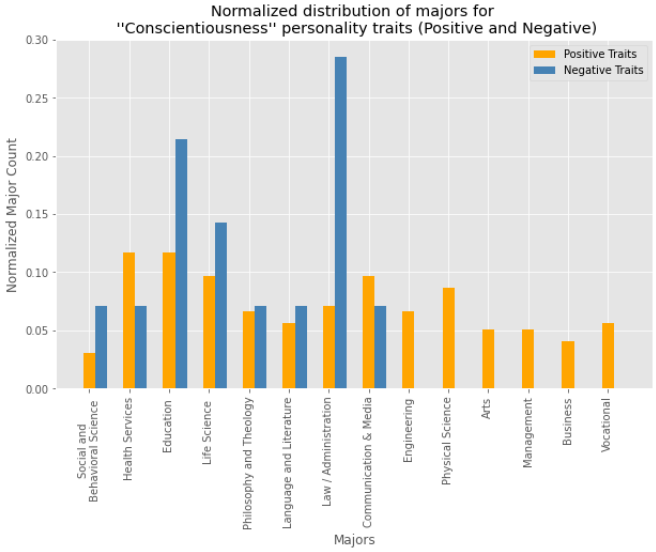

Fig. 1. Distribution of college majors based on Conscientiousness using results from the MTurk Survey

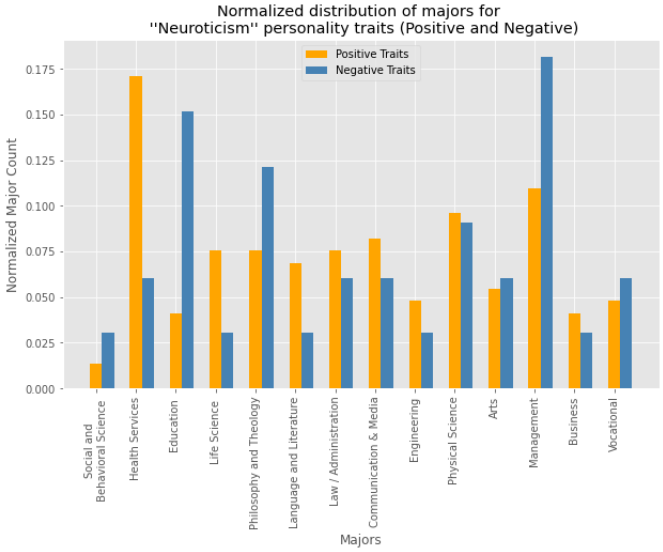

Fig. 2. Distribution of college majors based on Neuroticism using results from the MTurk Survey

accuracy using dimension reduction. Much of the time in machine learning, dimension reduction is a black box, and the feature sets derived from higher dimensions to lower dimensions have no intuitive real-world meaning. The OCEAN Index, however, reduces 10 TIPI questions to 5 interpretable attributes. From Figure 3, the decision tree performs well at a depth of 17 with an accuracy of 0.92 and later at depth 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.

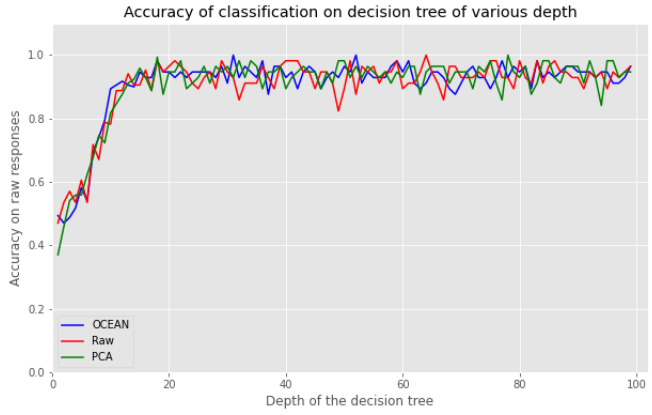

Fig. 3. Accuracy of the different decision tree over different tree depth

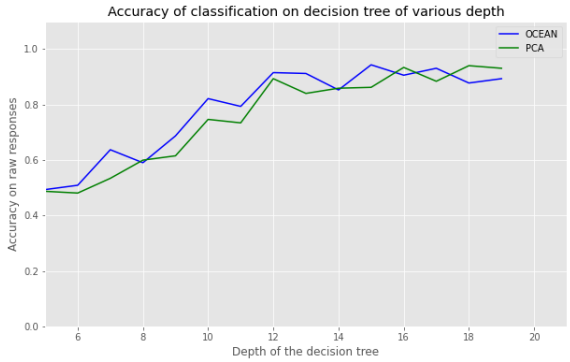

Fig. 4. Accuracy of OCEAN and PCA technique between depth 5-20.

C. Dimension Reduction by Principal Component Analysis

For the third technique after raw TIPI scores and OCEAN dimension reduction, Principal 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. The data was first fit into PCA to get the first five principal components. Scikit-learn was used for decision-tree classification. The decision tree is also affected by the depth of the tree - hence it is benchmarked with trees of depth 1 to 100 level. This allows for finding the optimal tree depth. Figure 4 shows the accuracy of classification over different depth of the decision tree.

D. Unique personality traits found in students preferring different majors

From the study, we were able to see the difference in different personality traits in students preferring each academic majors. For example, in figure 5, Business majors have the lowest "openness" score while philosophy and theology majors have the highest openness scores. This makes intuitive sense, as theology and philosophy majors are expected to deal with the wide range of views and people.

The distribution for openness and conscientiousness score for each major categories is presented in figure 5 and figure 6 respectively.

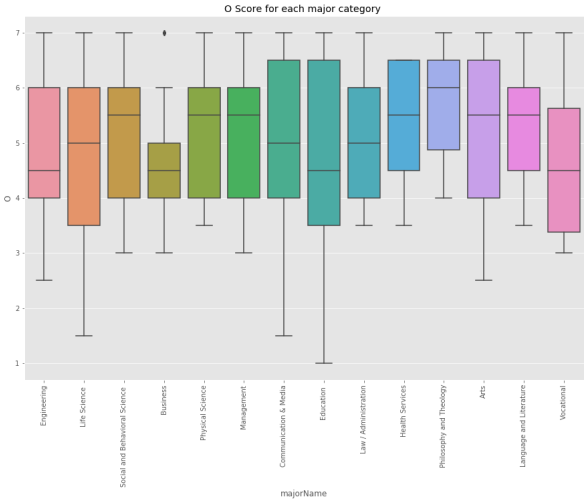

Fig. 5. Distribution of "Openness" traits among different academic majors.

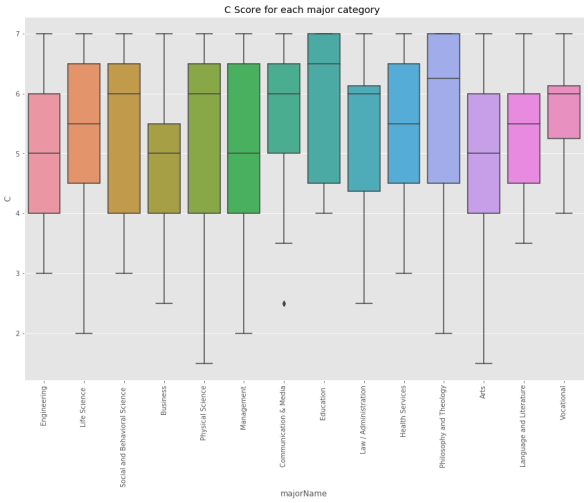

Fig. 6. Distribution of "Conscientiousness" traits among different academic majors.

V. Discussion

We began this research with three questions:

- R1: How effective is the use of personality traits in predicting a preferred college major?
- R2: How do expert-derived personality traits compare to a data-driven dimension reduction technique?
- R3: What are the unique personality traits found in students preferring different majors?

For this study, a framework to collect the user data and the survey was launched on Amazon's MTurk platform and also through in-university recruitment. Over 800 surveys were collected, and 543 were deemed valid and analyzed. The key contributions of this research are discussed below.

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

Based on both surveys, 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. 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 predictions can reinforce the existing students' notions of what they should be studying, our work can be viewed from a different perspective and used as a diagnostic and correcting tool. If we see that a large number of students are recommended certain majors because they answer personality traits question in a particular way, this can be used to reach out to them and organize programs that can introduce them to other career paths and choices.

B. Exploration of unique personality traits found in student preferring different majors

From the study, we were able to see the differences in different personality traits in students preferring each academic majors. For example, business majors have the lowest "openness" score while philosophy and theology have the highest openness score. This makes intuitive sense as well, as theology and philosophy majors are expected to deal with the wide range of views and people. Similarly, for traits like extraversion, vocational majors such as electrician or lab technician have the lowest score – meaning the most introverted while communication and mass media majors have the highest score. These scores can be use as diagnostic tool, for a student to focus on and develop certain skills that they need in their major and subsequently, career field.

C. Effect of dimension reduction technique in the recommendations

In this research, two different ways of dimension reduction were 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 data-derived Principal Component Analysis (PCA) technique. However, using OCEAN gives 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. 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.

VI. Future Work

There were some limitations and caveats in this project because of scope and availability of data. One portion of the data was heavily skewed to the white and female demographics, affecting generalizability. Similarly, because the short TIPI questionnaire was used in this survey, it can have biases on self-perceived traits and self-reporting. A future work with a broader data set and a larger questionnaire could help reduce these biases.

In this research, survey respondents were either young high school graduates or early college students. While surveying them gives a good insight in interest level on certain college majors, it does not prove success in their career field. Future work in the area could be a study that covers professionals in different fields and different majors. This would give a much clearer picture of career success in different fields for people of different personality traits.

References

[1] 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.

[2] N. R. Center, "National six-year and eight-year college completion rates."

[3] A. S. C. Association *et al.*, "State-by-state student to counselor ratio report: 10 years trends," 2015.

[4] A. Ghimire, Private Communication, 2019.

[5] L. J. Schneider and T. D. Overton, "Holland personality types and academic achievement." *Journal of Counseling Psychology*, vol. 30, no. 2, p. 287, 1983.

[6] 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.

[7] A. Allred, M. Granger, and T. Hogstrom, "The relationship between academic major, personality type, and stress in college students," *Eukaryon, 9*, 2013.

[8] 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.

[9] O. P. John, E. M. Donahue, and R. L. Kentle, "Big five inventory," *Journal of Personality and Social Psychology*, 1991.

[10] 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.

[11] D. of Education, "Collegescorecard.ed.gov. 2020. college scorecard data," 2020. [Online]. Available: https://collegescorecard.ed.gov/data/

[12] U. S. C. Bureau, "Retrieved november 20, 2021, from united states census bureau: https://www./quickfacts/fact/table/us/pst045219."

[13] 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.