Motivations/Influences for selecting a major for university students in the US south east region

N. Hoang* J. Anguiano† J. Mativo‡ U. George§

Abstract

In this paper, we examine student participant backgrounds and their motivations for choosing academic majors and their intents after graduation. Through descriptive statistics, we try to understand the main characteristics of the data set. Furthermore, we use the learning method Random Forest to determine if there exists particular grouping of the data that provide meaningful insight as to what influences a student's choice of major. We are mostly interested in discovering if there are any groupings that arise from the gender of the participants and if the field of study of the immediate family members has a significant influence. Findings suggest that most participants had developed interest with their majors before pursuing them at the university.

1 Main Text

Scholars (1 – 5) have time and again struggled in determining whether a formula or model can be established for students to use when faced with opportunities to select an academic major. Scholars show that the educational environment and experience exerts an effect on students' choice of majors. However, motivation to pursue specific majors in the Arts and sciences has minimal literature input. We explored participant background such as schools attended, extracurricular done, and immediate family education backgrounds to find if any relationship existed. We believe that if a strong relationship is found that leads to selection of majors, then resources can be used well to promote education endeavors to students and their respective families.

2 Theoretical Framework

Many factors contribute to a participants' choice of an academic major. The theory that seems to captures best for the motivation of a student in selection of a major is the Self-Determination Theory (SDT). The theory discusses intrinsic and extrinsic sources of motivation. The two sources have different triggers (6-7). Students who are indecisive the maybe counseled by using the SDT. Intrinsic motivation is referred to as undertaking action because it is interesting. Outcomes such as satisfaction, enjoyment, or happiness brought by overcoming a challenge is associated with intrinsic motivation. On the other hand, extrinsic motivation refers to action undertaken because it leads to a separable outcome such as family approval, praise from society, any valued action by others as externally prompted deed.

*San Diego State University
†San Diego State University
‡University of Georgia
§San Diego State University Department of Mathematics and Statistics, ugeorge@sdsu.edu

3 Exploratory data analysis

3.1 Demographics

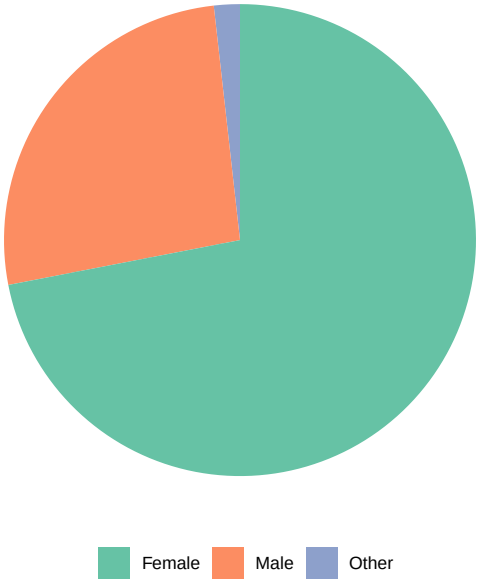

Figure 1: Gender of the participants

Participants in our study composed of 71.9% female, 26.3% male and 1.8% chose other. Fig 1 shows the distribution.

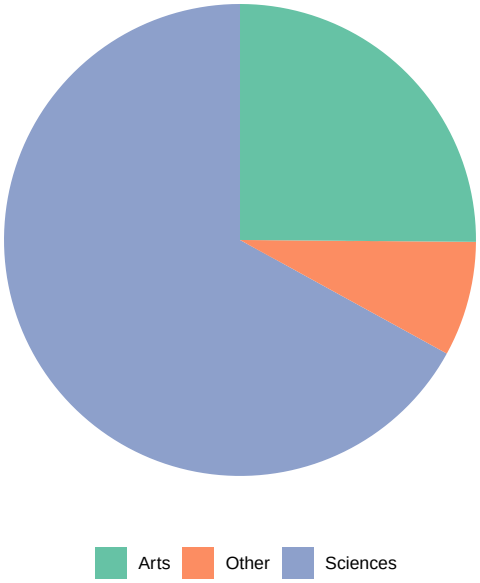

Figure 2: Major selection

Major selection is shown in figure 2. 25.1% picked the majors in the Arts, 65% chose Science, while 10% selected other.

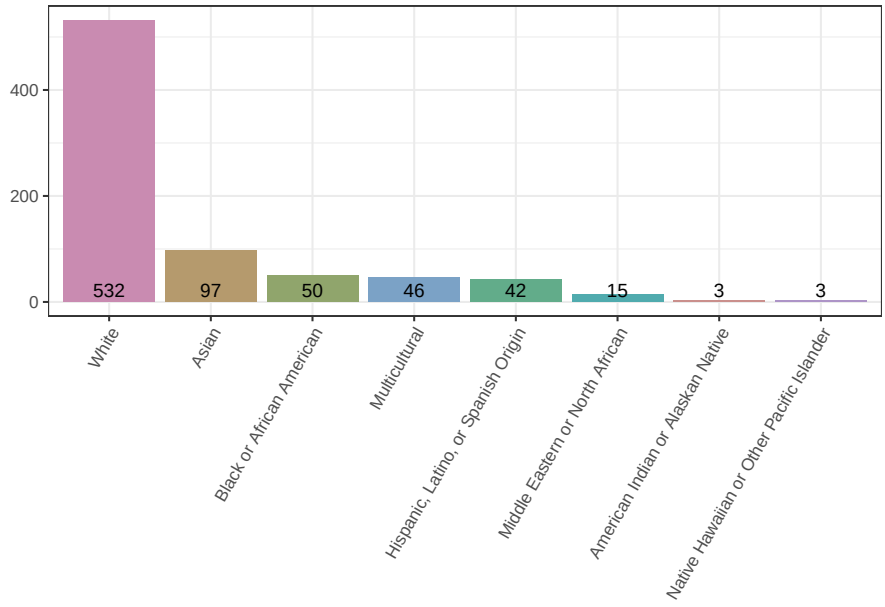

Figure 3: Ethnicity

Participants identified their ethnic grouping as shown in figure 3. Most of the participants were white 67.5%, followed by Asian 12.4%, then African American or Black 6.3%, Multicultural were 5.8%, Hispanics were 5.3%, Middle Easterns or North Africans were 1.9% while both American Indian or Alaskan Natives and Pacific Islanders represented 0.4%.

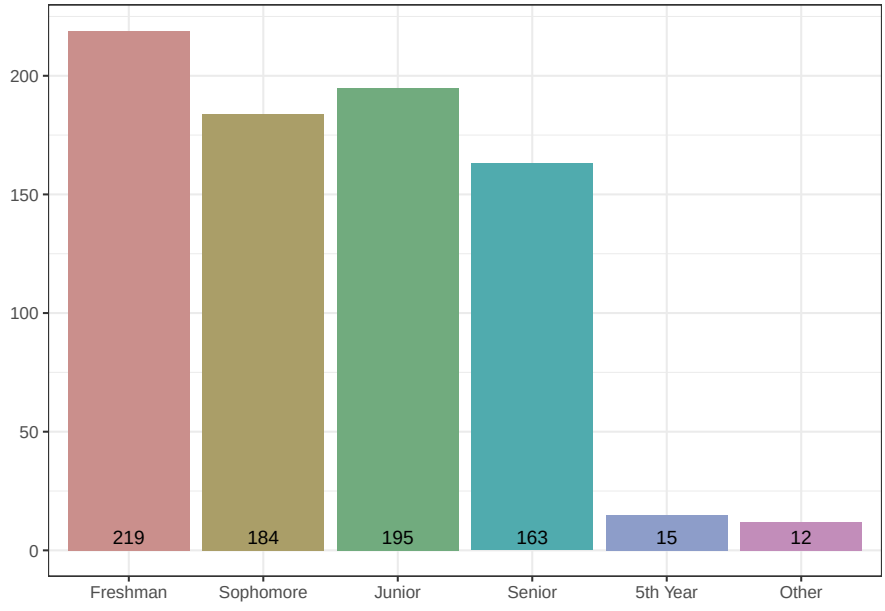

Figure 4: Participants college standing

Participants were represented almost equally with respect to their college standing. Freshmen had the highest representation at 27.8%, Sophomores at 23.4%, Juniors at 24.7%, Seniors at 20.7%, 5th year at 1.9% and other had a 1.5% representation. See figure 4.

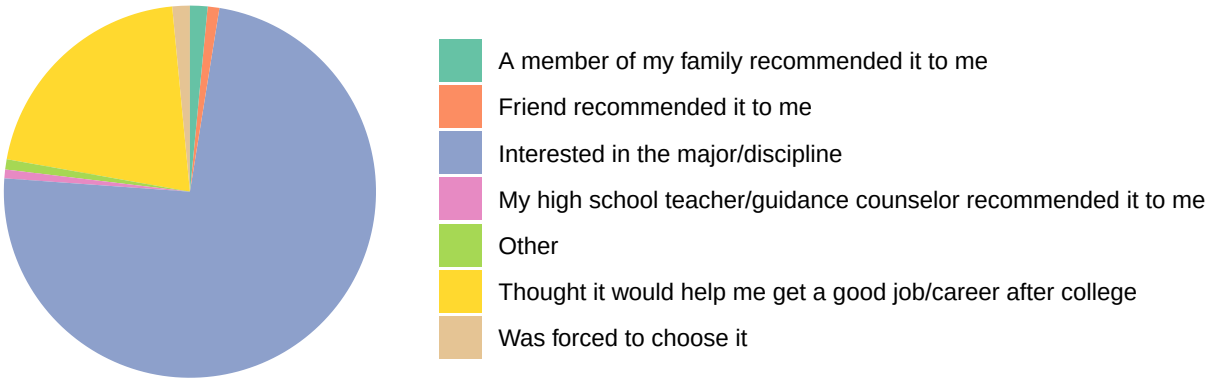

Figure 5: Most important reason for enrolling in the declared major

Two reasons that stood out as to why a participant chose a major were (a) thought it would help in getting a good job or career after college [20.7%], and (b) they were interested in the major or discipline [73.6%]. Other reasons included family member recommendation [1.5%], forced to choose it [1.5%], recommended by a friend [1%], recommended by high school teacher/counselor [0.8%], and 0.9% choose their response as other.

3.2 Family education background

The study ensued to determine any selection of major relationships between family members and participants. Results are shown in figures 6 - 10.

Looking overall, only 13.7% of the participants have family members that graduated with similar majors.

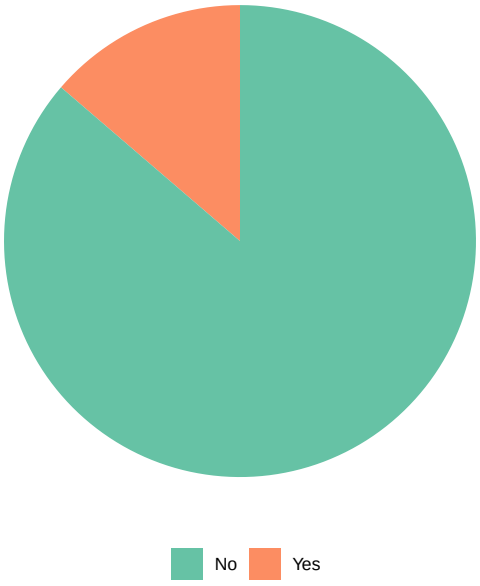

Figure 6: Proportion of family member with similar major

Of the families with members degrees, those who had similar degree majors as participants included 35.5% mother; 23.4% father; 11.2% both mother and father ; 2.8% grandfather; 1.9 grandmother ; 5.6% brother; 4.7% sister, and 14.9% relatives.

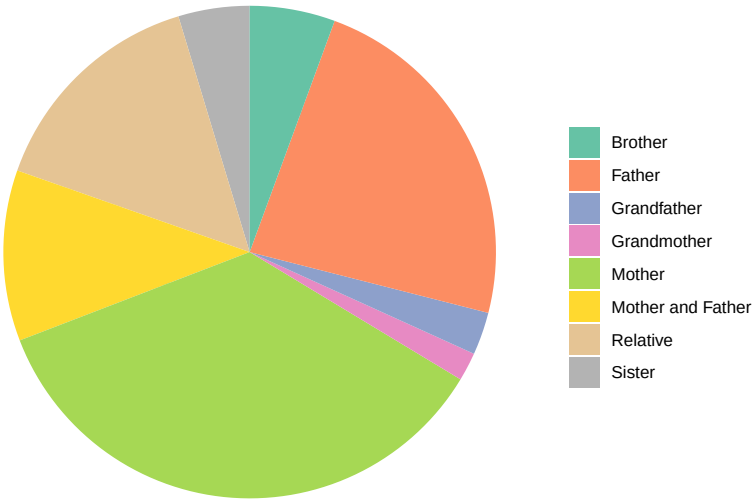

Figure 7: Family and participant similar majors

Half of those that declared science as their major had a family member with a degree in science.

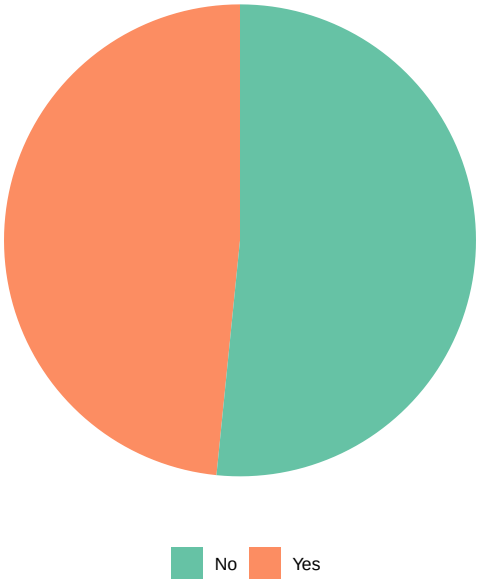

Figure 8: Participants with family members with majors in Science

Participants in the Arts major had fewer family members that had taken a major in the Arts.

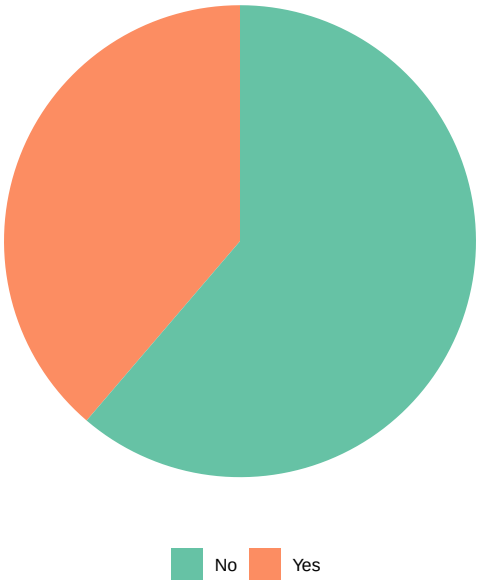

Figure 9: Participants with family members with majors in Arts

Engineering is viewed as an integrated degree for Mathematics, Science, and Technology. We set out to find whether participants had any relatives that had engineering degrees. Results indicate that 21% [165] stated that they had family members with a degree in Engineering while 79% [623] did not. When further asked who in their family had an engineering degree, results presented in figure 9 indicate that 52.7% were fathers, 6.7% mothers, 9.1% both mother and father, 4.3 % brothers, 3 % sisters, and relatives composing 24.2%, of whom half were identified as a grandparent.

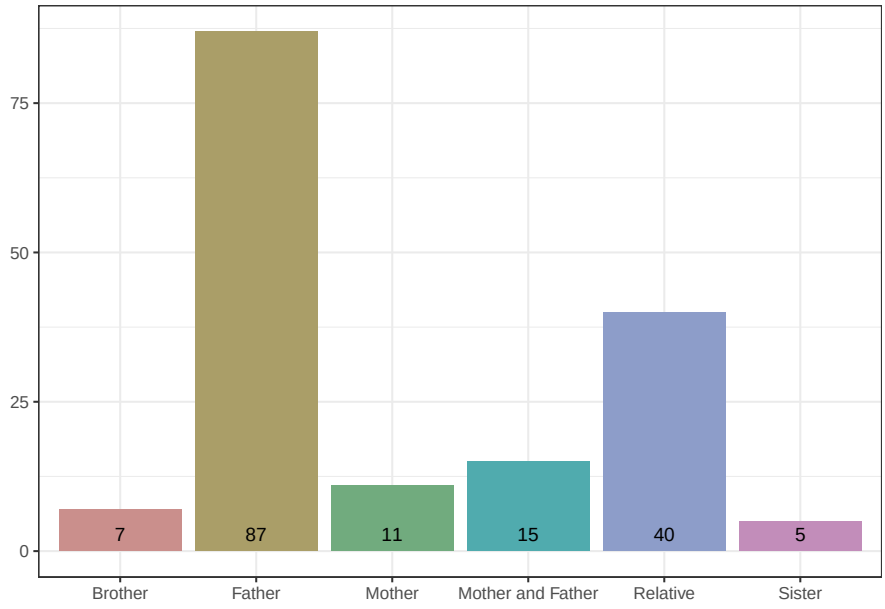

Figure 10: Family members with a degree in engineering

In seeking to understand family education background and how it might influence student major choice, we sought to find educational levels of participant parents' educational level. Results are shown in figure 11 and 12.

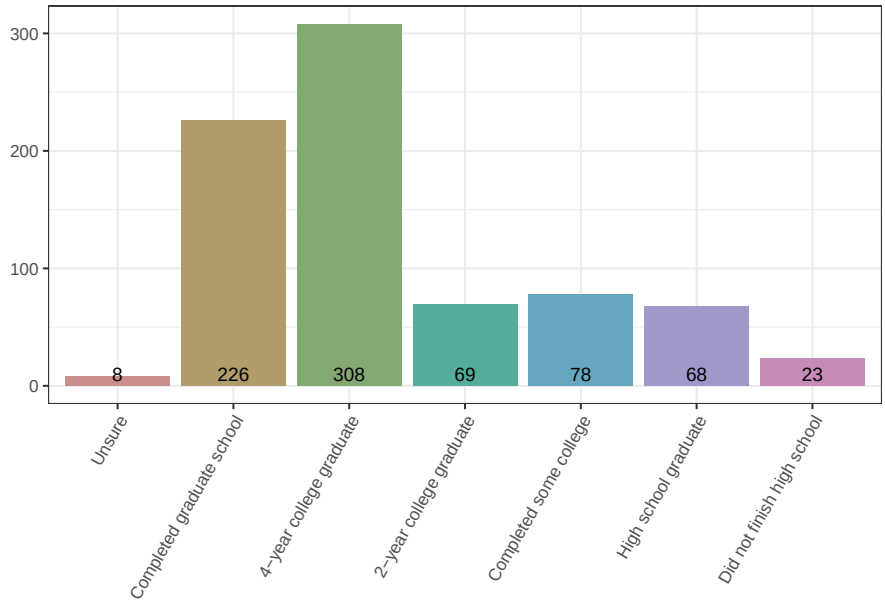

Figure 11: Educational level of the mother

Most mothers of the participants had either a 4-year college degree (39.5%) or had completed graduate school (29%). The Educational level of the fathers was similar to the mothers. Figures 11 and 12 shed similarities. The fathers 4-year degree represented a 35.3% while the graduate school was 33.3%.

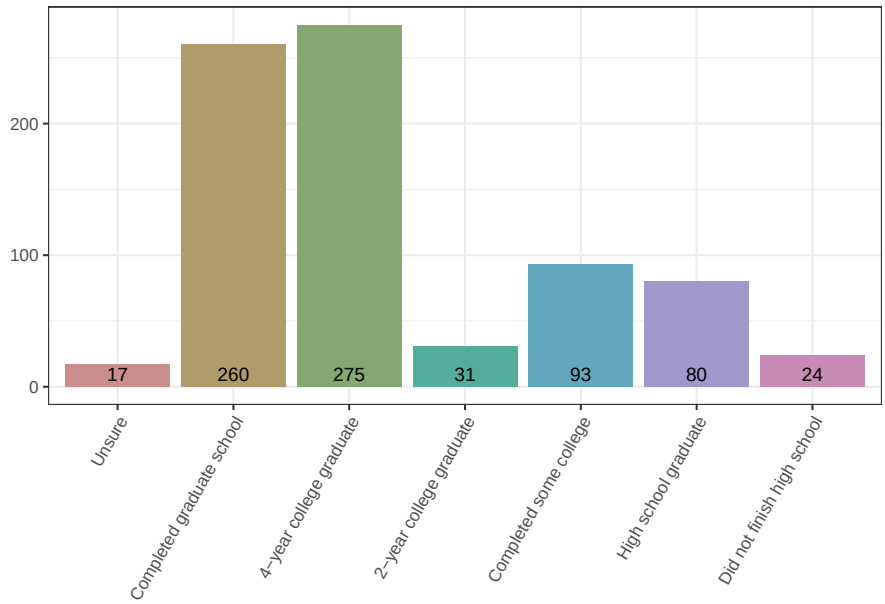

Figure 12: Educational level of the father

3.3 History of schooling and plans for after college graduation

Participants mostly attended public high schools (76%), public charter (7%), private non-religious affiliated (6%), private religiously affiliated (9%), and other (2%). Interestingly, only 22% (170) of participants attend a STEM related summer camp program, while 78% (596) did not. Of the participants, 33% (253) attended

an Arts related summer camp while 67% (511) did not. Now in college, most participants believe they are performing well in their classes as shown in table 1.

Table 1: Grade Point Average (GPA) of Students

GPA	Students	Percentage
Between 1.0 and 1.5	2	0.3
Between 1.6 and 2.0	0	0.0
Between 2.1 and 2.5	11	1.4
Between 2.6 and 3.0	61	8.0
Between 3.1 and 3.5	251	32.8
Between 3.6 and 4.0	440	57.5

Participants shared on endeavors they plan to pursue as presented in figure 13. As can be seen, over 50% of participants plan to pursue graduate school full time. An additional 13% will pursue work full time and graduate studies part time. An indication that they value education at a higher degree. 23.6% will work full time following graduation.

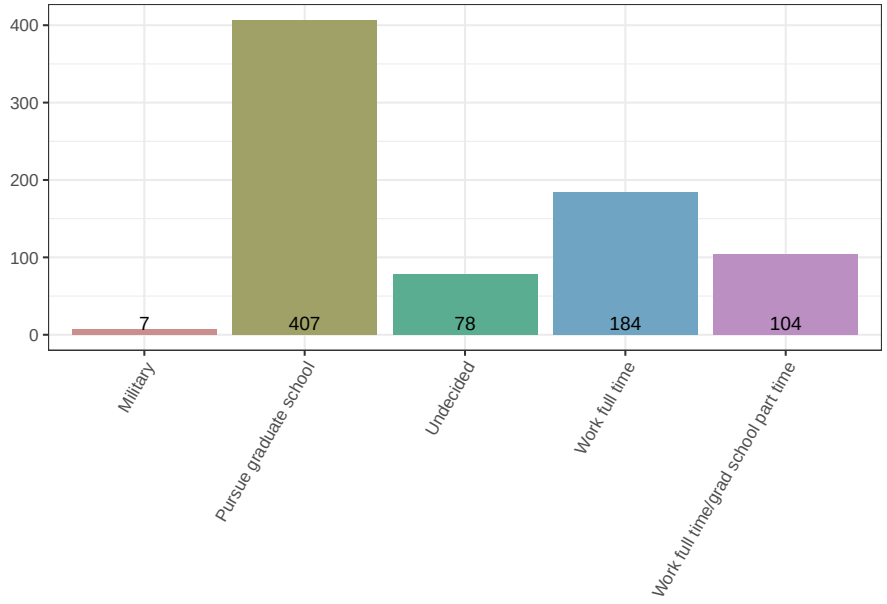

Figure 13: Endeavors after undergraduate work

We also look into how participants in each ethnicity choose their area of major. Table 2 records the number of students in each group. Given that there is only a small number of American Natives or Pacific Islanders so we can't really infer any information from them. However, looking at the Asian group, about 82.5% of them choose Sciences majors, the highest percentage overall, suggesting that the group is more into Science-related major than any other group. The group with the lowest percentage is Hispanic, Latino, or Spanish Origin, with only 59.5% choosing Science majors. As for Art, Black or African American has 32% choosing a major in the area, about nearly a third of the ethnicity's population.

Table 2: Distribution of participants' ethnicity in choosing major area

	Arts	Other	Sciences
American Indian or Alaskan Native	0	1	2
Asian	12	5	80
Black or African American	16	4	30
Hispanic, Latino, or Spanish Origin	13	4	25
Middle Eastern or North African	2	0	13
Multicultural	11	4	31
Native Hawaiian or Other Pacific Islander	1	0	2
White	143	44	345

4 Method

4.1 Theory

To start with, as a variation of tree-based method, random forest has been one of the most popular model to choose. In random forest, instead of building just one decision tree, we grow a bunch of them and when building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors [8]. Particularly, at the split, only a handful of m $\approx \sqrt{p}$ predictors to be consider. If the dataset has a really strong predictor, this will give more chance to see a pattern in other predictors than being shadowed by the really strong predictor. That's the whole idea behind random forest but before we can grow a forest, we must know how to grow a tree first [8]:

1. We divide the predictor space—that is, the set of possible values for $X_1, X_2, ..., X_3$—into J distinct and non-overlapping regions, $R_1, R_2, ..., R_J$. Or what we want to find are regions $R_1, R_2, ..., R_J$ that minimize the function RSS given by $\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$.
2. For every observation that falls into the region R_j , we make the same prediction, which is simply the mean of the response values for the training observations in R_j.

However, this might leads to potential overfitting the data due to the tree's complexity, whichs results in poor test performance. As a balance to the variance-bias trade, a strategy with smaller tree with fewer splits is proposed by growing a very large tree at first then reduces it down to a smaller tree that leads to the

lowest test error rate. According to [8], the algorithm to build a tree is:

Algorithm 1: *Building a Regression Tree* [8]

1. Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.
2. Apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of α.
3. Use K-fold cross-validation to choose α. That is, divide the training observations into K folds. For each k = 1,$\cdots$, K:
   (a) Repeat Steps 1 and 2 on all but the kth fold of the training data.
   (b) Evaluate the mean squared prediction error on the data in the left-out kth fold, as a function of α.
4. Return the subtree from Step 2 that corresponds to the chosen value of α.
   Average the results for each value of α, and pick α to minimize the average error.

Since our dataset is closely related to a classification problem so instead of the mean response of the training observations that belong to the same terminal node for the predicted response as above, it is the most commonly occuring class of training observations and also the class proportions among the training observations that fall into that region. Another difference is that we can't use RSS as a criterion for making the tree splits so we use the classification error rate. According to [8], there are two main ways to measure the classification error rate:

- The Gini index:
$$G = \sum_{k=1}^{K} \hat{p_{mk}}(1 - \hat{p_{mk}})$$
- The entropy:
$$D = -\sum_{k=1}^{K} \hat{p_{mk}} \log \hat{p_{mk}}$$

4.2 Illustration

Since it can be quite hard to visualize a whole random forest, which consists of a lot of trees, we first need to visualize a tree. Using the dataset ptitatic, which is the Titanic data of the passengers, we set up an illustration for a decision tree:

```
## pdf
##   2
```

Based on the figure, from the top-down, we can see the pattern from the dataset is that the probability of survival of a passenger is 38%. It is observed that if the passenger is a female (about 36% of the whole population), the chance of survival increases to 73%. However, if you are a male, the chance is only 19% and if the age is more than 9.5, it decreases by 2%. However, if the age is less than 9.5 then the survival probability is 53% and if the passenger has 3 or more siblings/spouses aboard, the rate jumps to 89% but if it is the other way then the chance diminishes down to 5%.

5 Results

5.1 The Random Forests

The initial random forests provided some insight as to what questions were most important in determining the groups in question. We began with the question of gender.

Figure 15 is used to make sure that enough decision trees are used to build the forest. The goal here is to see the plot stabilize as the number of trees increases. The results for the random forest are shown below.

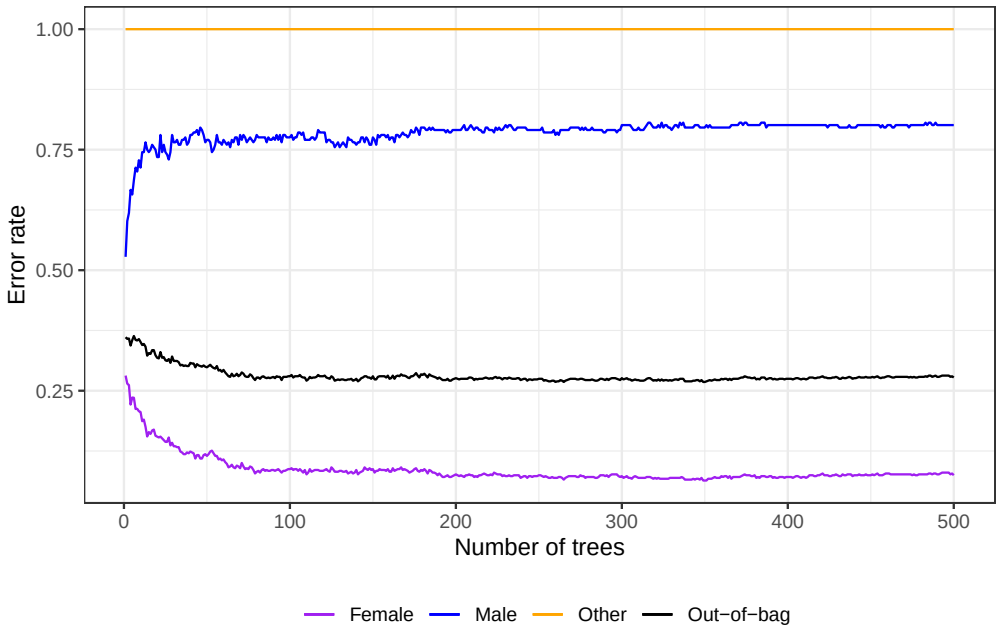

Figure 14: Error rate of classification based on the number of decisison tree used to build the forest for gender

Table 3: Confusion matrix with OOB estimate of error rate of 27.74%

	Female	Male	Other	class.error
Female	508	41	0	0.07
Male	157	39	0	0.80
Other	11	1	0	1.00

From the information above we can see that there were 500 trees created for this forest. The error in determining gender in general was at 27.74% but most of the error came from predicting males and others. The forest was significantly better at predicting females with an accuracy of about 92.6% versus an accuracy of about 19.9% for predicting males while there aren't any correct classifications of other. From figure 16, we can see that the participants' declared major and data reference from the system were two most significant factors that the random forest used to determine a student's gender. It worth noting that one known disadvantage of random forests is that they tend to give a higher importance to categories with more than 25 levels. In this case, both the DeclaredMajor and DataReference sections had more than 40 levels so here we might want to also consider other categorization of importance like major field area or their family influence of having similar major field area.

It turned out that it was much easier to determine if a student is a female using random forest. And the most important factors were student major and the students' field of study. It appeared that most of the female students were not simply studying in the field of arts but specific majors within the arts and/or specific majors within the sciences. This may explain why the random forest was mostly wrong in determining males who were well distributed across all majors and thus based on the survey R couldn't distinguish them based on their major alone.

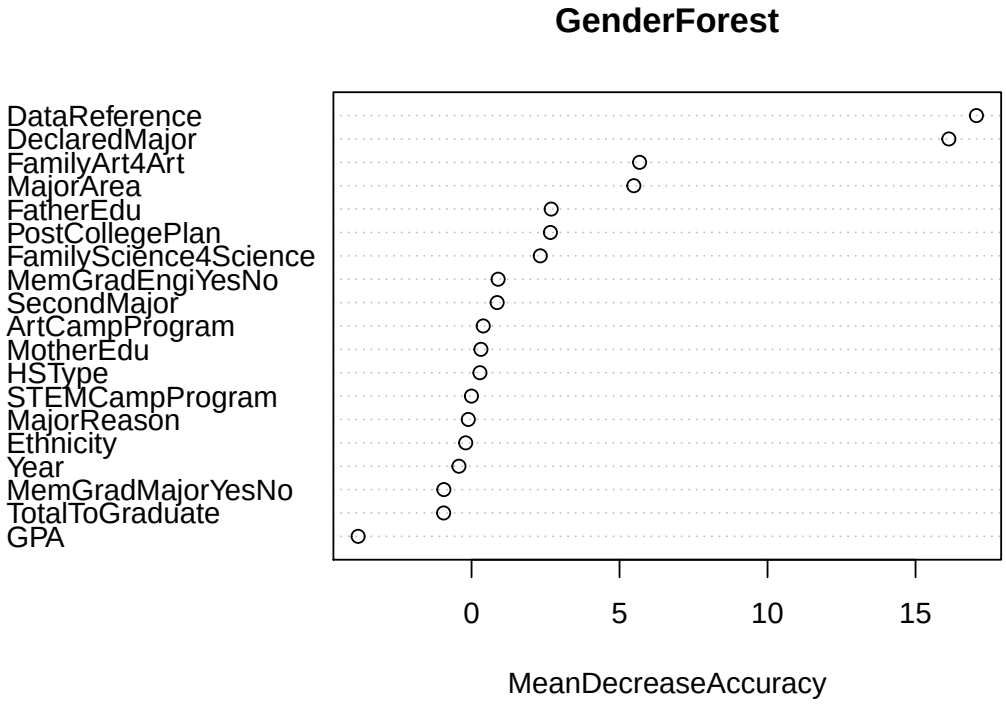

Figure 15: Factors that the random forest used to determine a student's gender

The students' field of study was next to be determined. The FieldForest was a Random Forest to predict the area of study. There were three possibilities for the area of study; Arts, Sciences or Other. The first step after creating the FieldForest was to plot the forest to determine if enough trees where created. As can be seen in the graph below, the Forest stabilizes at the very end which indicates that 500 trees were enough.

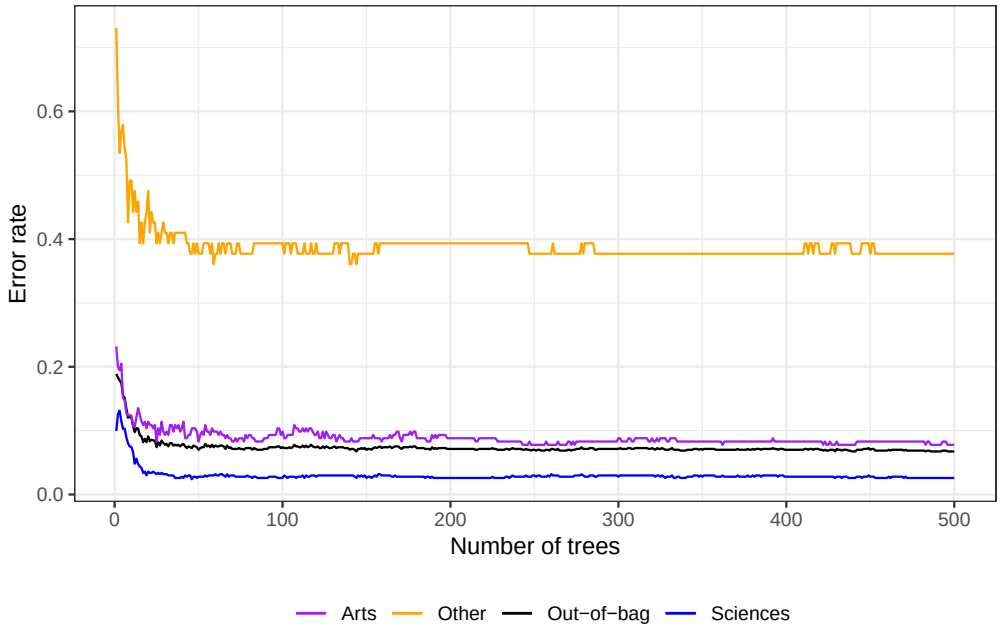

Figure 16: Error rate of classification based on the number of decisison tree used to build the forest for field of study

Table 4: Confusion matrix with OOB estimate of error rate of 6.74%

	Arts	Other	Sciences	class.error
Arts	178	2	13	0.08
Other	4	38	19	0.38
Sciences	7	6	490	0.03

The random forest was significantly better at predicting the field of study of the student given the data. Some may be concerned that the DeclaredMajor information is too similar to the variable of interest, but the random forest is not aware of that and is able to determine that on its own. Also, even while omitting the declared major data, neither the error nor variable importance plot changed significantly so we opted to leave that information in.

The most important variables were the students' family. So according to the random forest if the student had family in the same field of study then it is most likely that the student remained in the same field. Also, it appears that there is a distinction between the arts and sciences and why students chose their specific majors. Lastly, students' gender was not of particular importance in determining their field of study; the influence of their family and their motive for choosing the major was much more important.

Originally, we intended to also investigate the influence of family members and this forest confirms that family has a significant influence for students selecting their field of study. Additionally, from the data, about 39% of students in the arts had family members in the arts and about 48% of students in sciences had family members in the sciences.

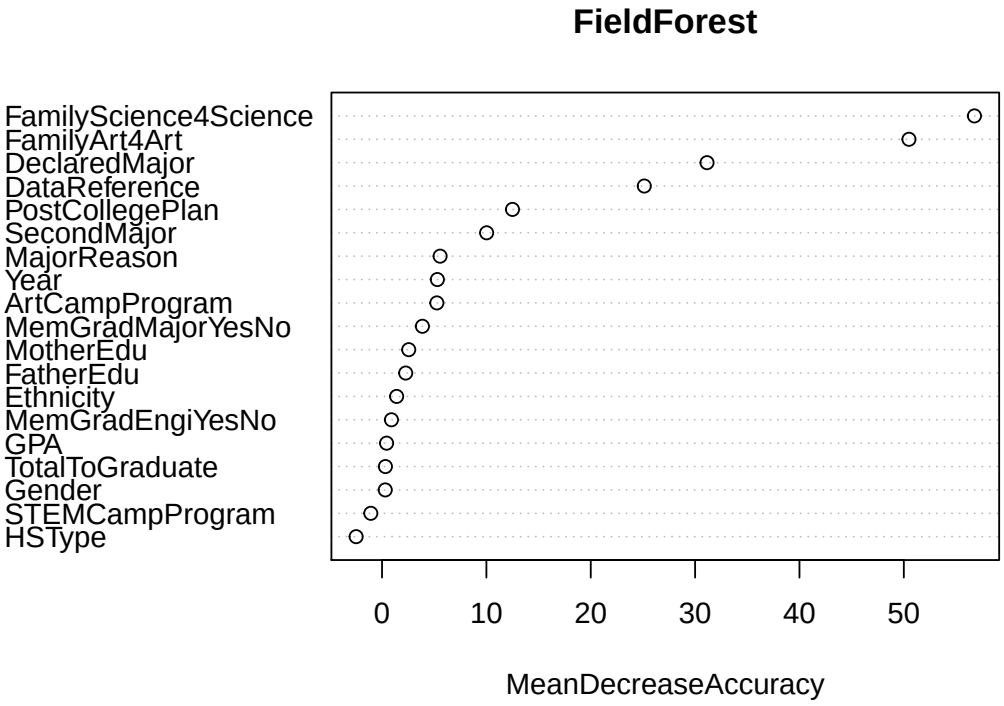

Figure 17: Factors that the random forest used to determine a student's field of study

5.2 Random forest clustering

Before getting into the analysis of the method, there are a couple of things that we first need to know about it. Random forest is not an actual clustering technique per se but can be used to create distance metrics based on proximity that input into traditional clustering methods, in this case, we will use Partitioning Around Medoids or PAM. The proximity represents the percentage of trees where the two observations appear in the same leaf node. Since our dataset is dealing with purely categorical data, it is hard to use the traditional distance metrics like Euclidean, Mahattan and so on, proximity seems like a good option here. However, for large problems, even using a sparse matrix, the very nature of the approach causes the proximity matrix to get huge so this might be the reason why unsupervised approach of random forest might not be favored. Another thing to notice is that we choose PAM, which is similar to k-means but uses the median, because we believe that the method is more robust to outlier and noise. The main goal of the unsuperised clustering method is an attempt to discover the trends that might exist in the dataset which we can't see normally.

First, in order to choose the optimal number of cluster for the method, it is best to use the average silhouette plot based on the proximity matrix produced by unsupervised random forest over a range of the number of clusters k. The idea is that the average silhouette indicates a good clustering, the higher the width, the better it is so the optimal number is the one that maximize the width. According to figure 19, the optimal number of clusters is 3.

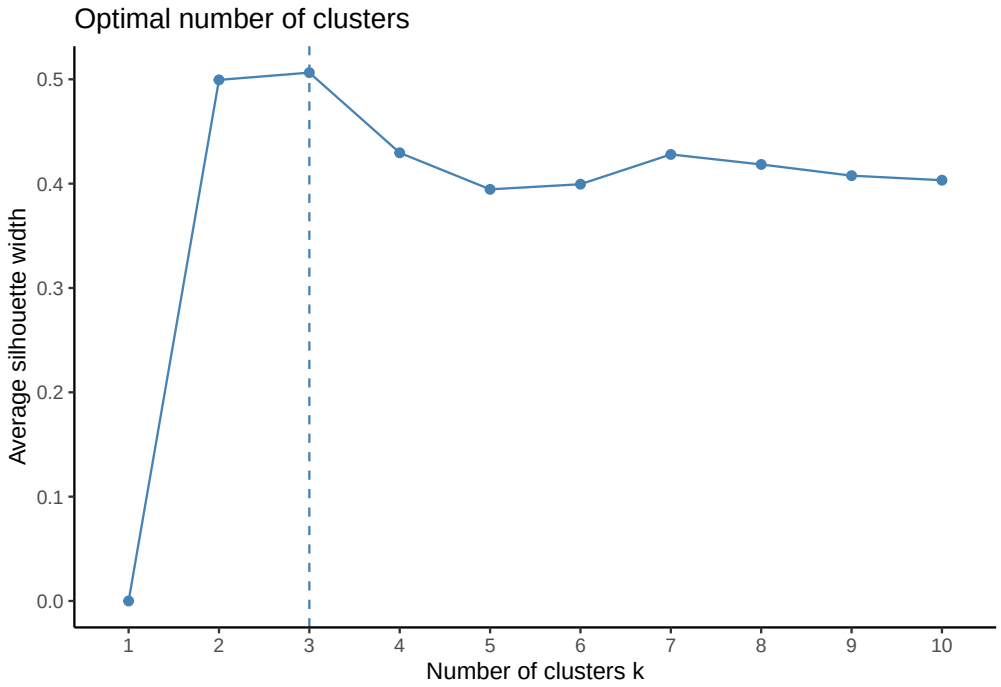

Figure 18: The average silhouette plot

In order to visualize the partitioning results, we draw out a scatter plot of data points colored by cluster numbers. Since the dataset contains more than 2 variables to predict, the Principal Component Analysis algorithm was used to reduce the dimensionality of the data so in figure 20, the first two principal dimensions are used to plot the data.

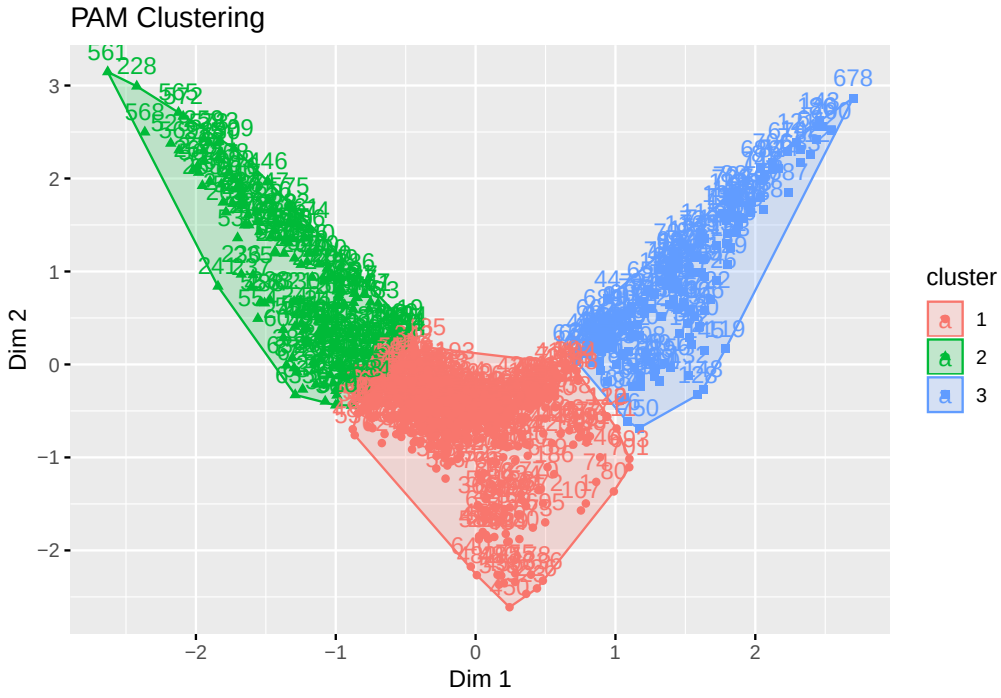

Figure 19: The multi-dimensional scaling plot of PAM clusters

Figure 20 shows the seperation between the clusters. However, it seems that cluster 1 is closely near both cluster 2 and 3, indicating a certain mix that is not too far off from both of them. While cluster 2 and 3 might represent similar characteristics due to the same range of values on dimension 2 but posess different identities based on dimension 1. To look further into the attributes of the clusters, we go into the analysis of the frequencies of each cluster in each variable along with figure 21 and found that:

- Cluster 1 has the most student, nearly 52% of the whole dataset, with all-rounded characteristics of every variable, suggesting no interesting information in any particular pattern.
- However, it's different for cluster 2 and 3. While cluster 2 has 27% of the population, all of the students in this group are science-oriented major, mostly focused on biology area. On the other hand, the rest of 21% of the dataset is in cluster 3, with the majority of 94% of the students are art-related majors and 6% is in other major area.
- The majority of students in cluster 3, up to 88%, chooses the major reason to choose their major is because they are interested in the art discipline itself. At the same time, even though the majority of cluster 2's reason is interest in the major, there is a pattern that nearly 36% thought that the majors they choose would help them get a good job or career after college.
- Looking into the mother education backgroup, most of the mothers in both group 2 and group 3 finished 4-year college, with 38.5% and 44% respectively. Correspondingly, 32% of the mothers in group 2 and 28% of the mothers in group 3 completed graduate school.
- Considering father education background, most of the fathers, with 42%, of the students in group 2 completed graduate school with 27% finish a 4-year college. On the contrary, for cluster 3, the majority of 44% of the fathers are 4-year college graduate while 27% completed graduate school.
- Last but not least, most of the students in group 2, nearly 84%, want to pursue graduate school after finishing college while 8% want to work full time and finish graduate school part-time. However, the majority of 41% of the students in cluster 3 want to work full-time while nearly 30% want to pursue graduate school.

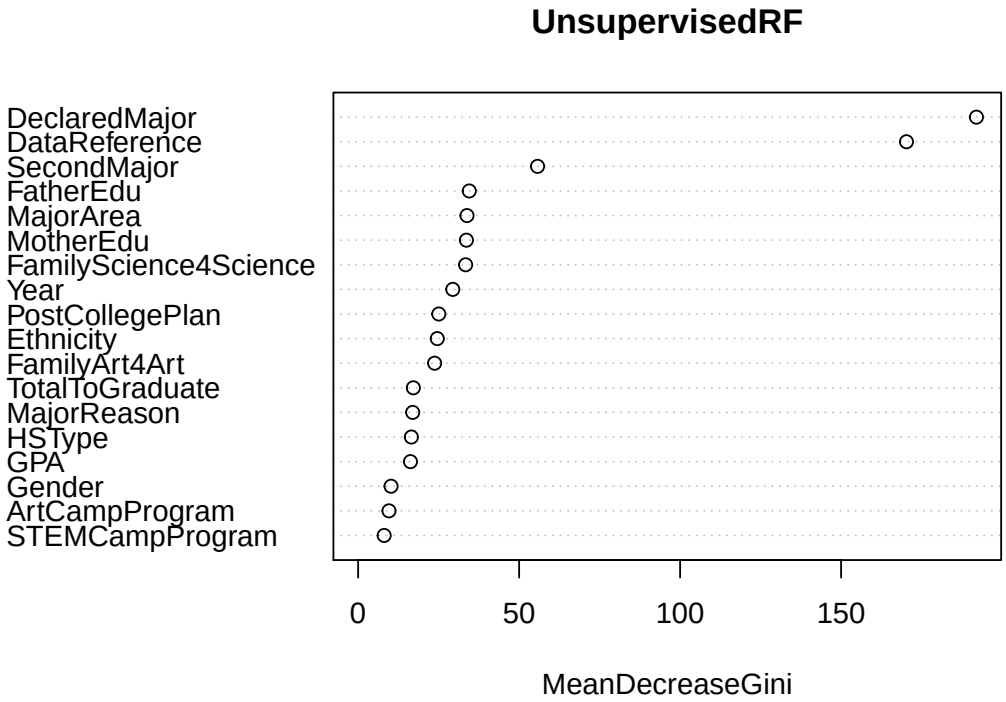

Figure 20: The variable importance plot of unsupervised random forest

6 Conclusion

In general, a students' major is mostly influenced by their attained interest up to the point where they must choose a major and mostly influenced by their family and not necessarily distinguished by their gender. From the responses to the survey about 74% of students chose their major because of their interest in the discipline and about 20% chose their major because they believed it help them get a job after college.

Some precautions to consider in regard to the results in this paper are that even though a person's gender was not the greatest determinant in their major decision, it may still be statistically significant. Also, ethnicity may also have played a role in the results of this survey since about 67% of students identified as white and 12.5% as Asian. The remaining six ethnicities completed the last 20.5% of the responses with no one taking more than 8.5% of the total. So perhaps people of other ethnicities are more greatly influenced by factors other than their family.

References

[1] Cebula, R. J. Lopes, J. (1982). Determinants of student choice of undergraduate major field. *American Educational Research Journal*, 19, 303-312

[2] Mauldin, S., Crain, J. L., Mounce, P. H. (200). The accounting principles instructors' influence on students' decision to major in accounting. *Journal of Education for Business*, 75(3), 142-148.

[3] Brown, S. D., Rector, C. C. (2008). Conceptualizing and diagnosing problems in career decision-making. In S. D. Brown, R. W. Lent (Eds.), *Handbook of counseling psychology* (4th ed., pp. 392–407). New York, NY: John Wiley.

[4] Carr, A., Rossier, J., Rosselet, J. G.,Massoudi, K., Bernaud, J., Ferrari, L., ... Roche,M. (2014). The career indecision profile: Measurement equivalence in two international samples. Journal of Career Assessment, 22, 123–137

[5] Mativo, J., Womble, M., Jones, K (2013). Engineering and technology students' perceptions of courses. *International Journal of Technology and Design Education.* 23(1), pp 103-115. https://doi.org/10.1007/s10798-011-9167-3

[6] Deci, E. L., Ryan, R. (1985). Intrinsic motivation and self-determination in human behavior, New York, NY: *Plenum*

[7] Deci, E. L., Ryan, R. M. (2000). The "what" and 'why" of goal pursuits: Human needs and self-determination behavior. Psychology Inquiry, 11, pp. 227-268

[8] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. An Introduction to Statistical Learning: with Applications in R. Springer Publishing Company, Incorporated.

[9] Andy Liaw, Mathew Weiner; Fortran original by Leo Breiman, and Adele Cutler. *randomForest: Breiman and Cutler's random forests for regression and classification*, 2014. R package, https://CRAN.R-project.org/package=randomForest.

[10] Terry Therneau and Beth Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2014. R package, https://CRAN.R-project.org/package=rpart.