Career choice and gender: an exploration through random forest classifiers and a picture of Spain's university students.

Valeria Orozco Castiblanco

Abstract

Career choice is influenced by a vast number of variables such as class, gender, school performance and participation in career development activities. Previous research has shown that career differences by gender might be due to cultural biases and gendered socialization. In the present paper, the relationship between career expectation and gender is explored across two main questions. The first one is related with the existing differences in gender over areas of study in Spanish universities, and the changes of such gendered differences across a span of 4 years. The second question revolves around finding the strongest variables related to the type of career a student decides to enroll in, to determine if gender seems relevant or not at the moment of prediction when implementing an ML model.
To answer the first question a statistical analysis was made on a dataset describing career choice and gender across the last 4 years in Spain, for the second one, a random forest(RF) algorithm was used with the purpose of examining if there is a relationship between gender and career choice. The results show that Spain follows a trend when it comes to career choice, with women outnumbering men in almost all areas of study but engineering and architecture. Similarly, the results from the implementation of RF show a strong relationship between gender and area of study, following results from previous research. Limitations and further recommendations are described in the discussion section.

Keywords: gender, career choice, Spain, random forests.

---

1. Introduction

Career choice is influenced by a vast number of variables. A recent publication of OECD explored the differences in career expectations of 15 to 16 year olds across the globe, and found that variables like class, gender, school performance and participation in career development activities were strongly related to career choice[1].

Among the variables that are known to affect career expectations, gender seems to be one of the most significant ones, even more than academic performance. Based on the results from the OECD report, teenagers (girls and boys) with the same learning outcomes in PISA are likely to choose a different career. Among those with high results in mathematics or science, boys, in average, were more inclined to show interest in engineering or science-related careers than girls[1].

The question of why there is a noticeable divergence in career choice by gender has been studied before and explanations tend to rely on the heavy influence of environmental factors

that surround students while growing up. For example, Konrad [2] suggests that the social environment where children grow shapes beliefs about “appropriate” behaviors according to someone’s gender. Along these lines, traditional gender roles postulate that young women have better social skills and that young men are good in mathematics and science. [3]

The most prevalent biases related to gender in education are related to student’s abilities and cultural fit. The first one refers to the belief that men have more ability, talent or potential for STEM (science, technology, engineering and math) and the latter is related to the type of person that is typically seen in the field, for STEM the stereotype goes around socially awkward and technologically oriented men. Which permeates not only the adult world, but the classroom. [2]

It has been found that girls that have had gender biased teachers, tend to be induced to underperform in math and report lower self confidence on their math ability, thus growing on heavily gendered stereotyped environment can create barriers that prevent girls from developing interest in STEM [4] as motivation and self confidence are key factors in influencing STEM education.

There has been increased awareness about the existent of e gender-related biases in education, thus promoting further research on how stereotypes permeate educational institutions, teacher instruction and student’s career choice. In spite of growing literature on the subject, there is still a marked difference in career choice by gender in recent student applications.

Some suggestions to reduce the gender gap on STEM environments are on one hand, building educational environments free from gender bias and on the other hand, promoting supporting networks where confidence is endorsed.[2]

The present work is inspired by the question about the existence of any differences between career choice by gender in the Spanish scenario.
A series of questions were posed as part of an effort to explore the data set on career selection.
Questions were the following:

1. Are there significant gender differences in career choice?
2. What are the careers with more noticeable differences in gender on students enrolled in Spanish universities?
3. Has there been any change in the career choice dynamic in the last three years in Spain?
4. What are the strongest variables related to the type of career a student decides to enroll in? and can gender be predicted based on demographic and academic information?

2. Method

With the aim of exploring the first question related to career choice by gender in Spain across the last 3 years, a statistical analysis of information from students entering their first year of university was made using age, gender, year of enrollment and area of study as main variables

The data was collected from the data catalogue opened by the Spain ministry of universities. (https://www.universidades.gob.es) The original dataset was later organized using various relevant variables for the replication of the study such as gender of students, age and career choice.

Information from the dataset comprises the following columns:

- *Age:*
  - 18-21
  - 22-25
  - 26-30
  - 30+
- *Gender:*
  - Female
  - Male
- *Quantity:*

  Represents the number of women or men under a certain period, enrolled in a certain career.
- *Period:*
  - 2019-2020
  - 2020-2021
  - 2021-2022
- *Career:*
  - Social sciences
  - Arts and humanities
  - Engineering and architecture
  - Health sciences
  - Sciences
  - Social sciences

A descriptive analysis of the data was performed with the aim of exploring changes from 2019 to 2021 in career distributions and differences in career choice by gender.

For the second question, a larger data set of enrollment from Open University was used.

Open University is a public British Online University. The data set describes general demographic information about the students, their academic performance and the area of study they have chosen. (https://analyse.kmi.open.ac.uk/open_dataset)

Information from the dataset comprises the following columns:

- *Gender*
  - Feminine or Masculine
- *Region*
  - Regions from the UK. (Scotland, London, Wales, Ireland)
- *Highest education*
  - Post graduate
  - A level: Beyond high school but not University
  - Non formal education
  - Lower than A level
  - HE level: Higher education
- *Imd band*
  - Index of Multiple Deprivation band of the place where the student lived during the module-presentation, going from 0 to 100%
- *Age band*
  - 0-35
  - 35-55
  - 55+
- *Studied credits*
  - Number of credits studied in the institution
- *Disability*
  - Yes or no
- *Final results*
  - Pass
  - Withdrawn
  - Fail
  - Distinction
- *Code module*
  - A module(area of study) on which the student is registered.
- *Number of previous attempts*
  - The number of times the student has attempted this module.

To answer the questions, what are the variables that mostly influence the type of career a student decides to enroll in? And can a student gender be accurately predicted based on demographic information and school performance? a random forest classifier model was used.
The model is suitable to solve the question since it tends to perform well on categorical data.

2.1 *Data preparation:*

Spanish Dataset:

The original data from the Spanish government had to be reduced to the information that concerned the study. For example, the original data set had information conveying career choice by gender since 2015, while also describing the number of students enrolled by style of education(online or presential) and institution(public or private). Since the data set was a statistical summary rather than a single value dataset, those variables were not taken into account. Aside from deleting unwanted columns, no changes were made to the dataset.

Open University dataset:

The dataset from Open University had to be restructured to fit a random forest algorithm since many of the variables were categorical.

On one hand, categorical variables such as region, final_result, highest_education, imd_band, age_band and code_module were converted into dummy variables, while gender and disability were converted to a numeric variable.

2.2 Missing values:

Because having missing values on the data set affects the outcomes of the model, those values that were found missing were filled. In the case of the data set, only the Imd band column was lacking information, missing values were filled by using an even distribution across all possible responses, 0 to 100% in 10 % ranges, and therefore the original distribution of the variable was not affected.

The training was made on 70% of the data and the remaining data was used for testing. After configuring train and test sets, the output for both training and testing set was resampled, more specifically the answer ‘F’ was oversampled to reach the same number of observations as ‘M’ using SMOTE.

2.3 Model used

To determine whether the person is female or male, based on the information provided in the data set, a random forest algorithm was used.

2.4 Evaluation metrics

To evaluate how well the model performs in the categorization problem a confusion matrix was used, accompanied by precision and recall metrics derived from the confusion matrix.

2.4.1 *Precision:*

$$precision = tp/(tp + fp)$$

Precision answers to the question of how many retrieved items are relevant.
Where *tp* refers to true positives, meaning the number of correct classification out of the students identified as female or male and *fp* refers to false positives, that is to say, the number of students identified within a category different from the correct one.

2.4.2 *Recall:*

Recall answers to the question of how many relevant items are retrieved from the whole set of each class.

The way to measure it is the following:

$$recall = tp/(tp + fn)$$

Where *tp* refers to true positives, and *fn* to false negatives, that is to say the number of observations from a class that were marked from a different class than the correct one from the whole set of positives.

To verify if indeed career choice or in this case, subject choice is relevant when categorizing gender, a feature importance measure of Mean Decrease in Impurity was developed in addition to the aforementioned metrics.

4. Results

Changes of career choice in Spain by gender from 2019-2021

The exploration from the dataset reveals some trends. Many of them aligned with the findings from the literature. For instance, the dynamic between men and women in terms of career choice are replicated, mainly for engineering and social sciences, as seen in figure 1. Nevertheless, in sciences women slightly outnumber men, in contradiction to what has been found in the literature.

It is interesting to notice how trends might be changing in the enrollment period of 2021, where women's presence in engineering augmented as well as men's enrollment in social sciences.

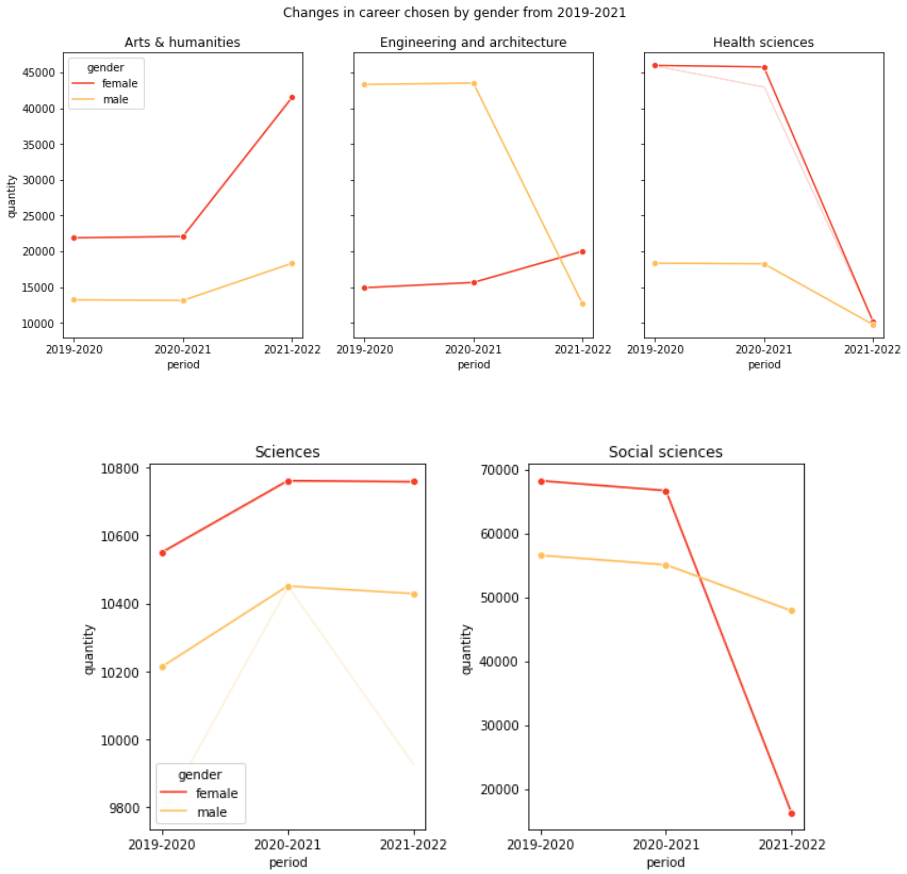

Figure 1: Number of women and men enrolled per year and area of study.

The reason for women outnumbering men in almost all areas of study is due more women enrolling to university than men in the entries from 2019 to 2021 as seen in figure 2.

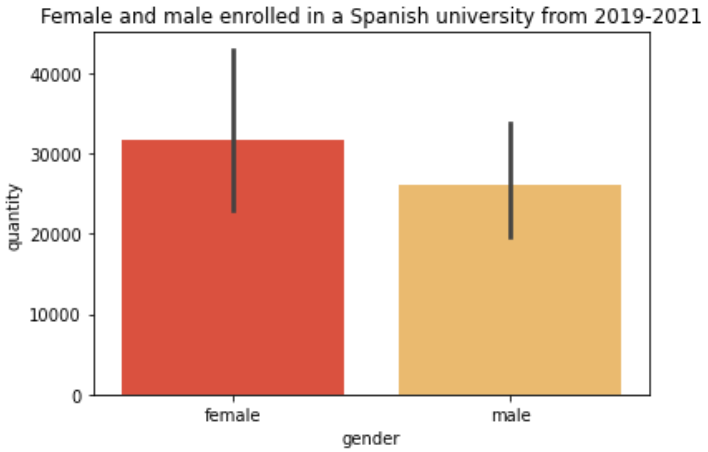

Figure 2: Female and male enrolled in a Spanish university from 2019-2021. Being women enrolled higher in number than men.

Regarding the age of those enrolled in a Spanish university, they are mostly between 18 to 25 years of age, at which the differences in career choice seem more pronounced. However, in older students, differences in gender seem to be more nuanced, as seen in figures 3 and 4.

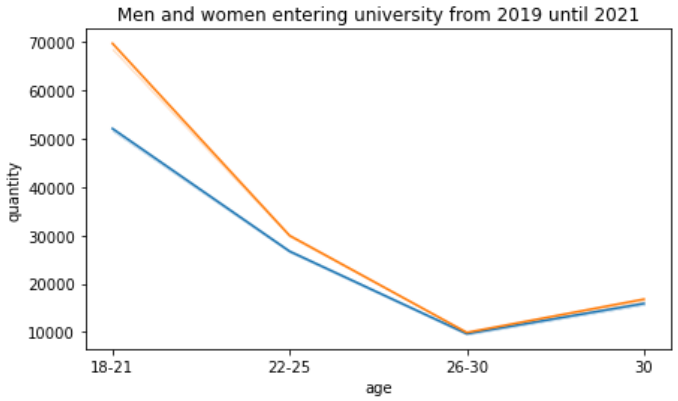

Figure 3: Ages of students enrolled in a Spanish university from 2019 to 2021. Most of the enrolled students, being aged 18 to 25 years.

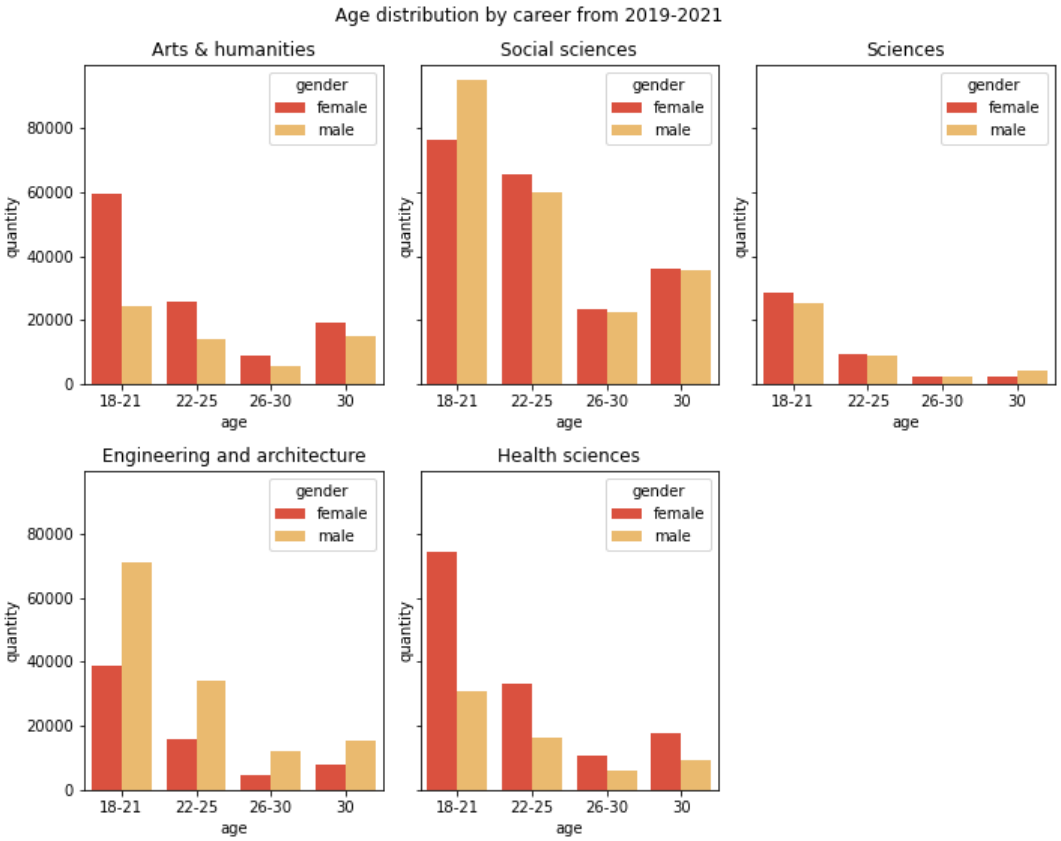

Figure 4: Age distribution of women and men across career and gender. Where differences in career choice are more noticeable in younger students, except for trends in engineering/ architecture and health sciences.

Finally, the analysis showed a decrease in enrollment from all students in the period of 2021-2022, especially for the younger students as seen in figure 5. This might have been due to COVID 19 restrictions and the economic impact it might have had on students and their families.

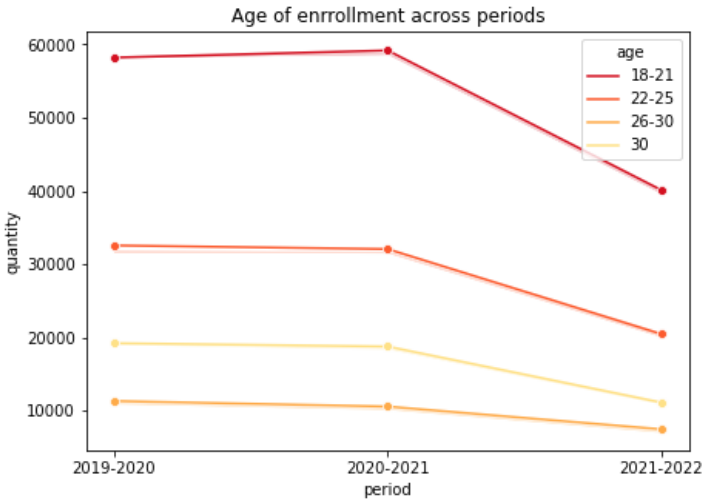

Figure 5: Age of enrollment across periods from 2019 to 2021. There is a visible trend showing a decreased enrollment from students, mainly from those from 18 to 25 years.

Prediction of gender based on demographics and course choice

When using Random forest to categorize gender the accuracy of the model reaches 70%, showing better performance for men than women, as shown in the figure below.

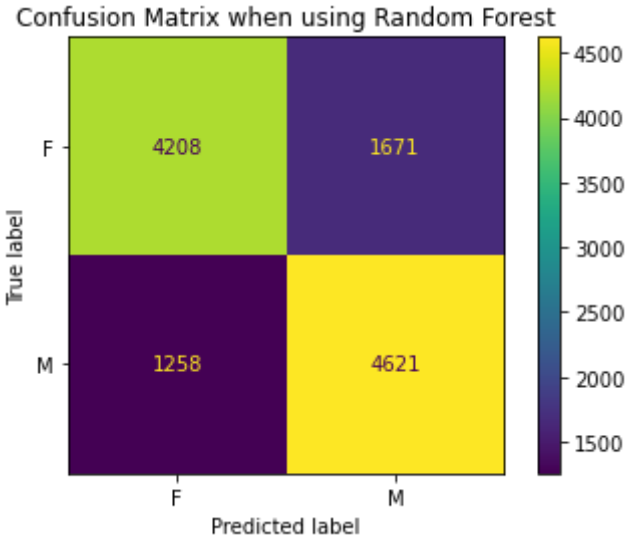

Figure 6: Confusion Matrix after using Random forest, 71% accuracy for female categorization while 78% accuracy on men categorization.

The features that were found to be more relevant were the courses, especially the code BBB and the number of credits a student completed on the student record, as seen in figure 7.

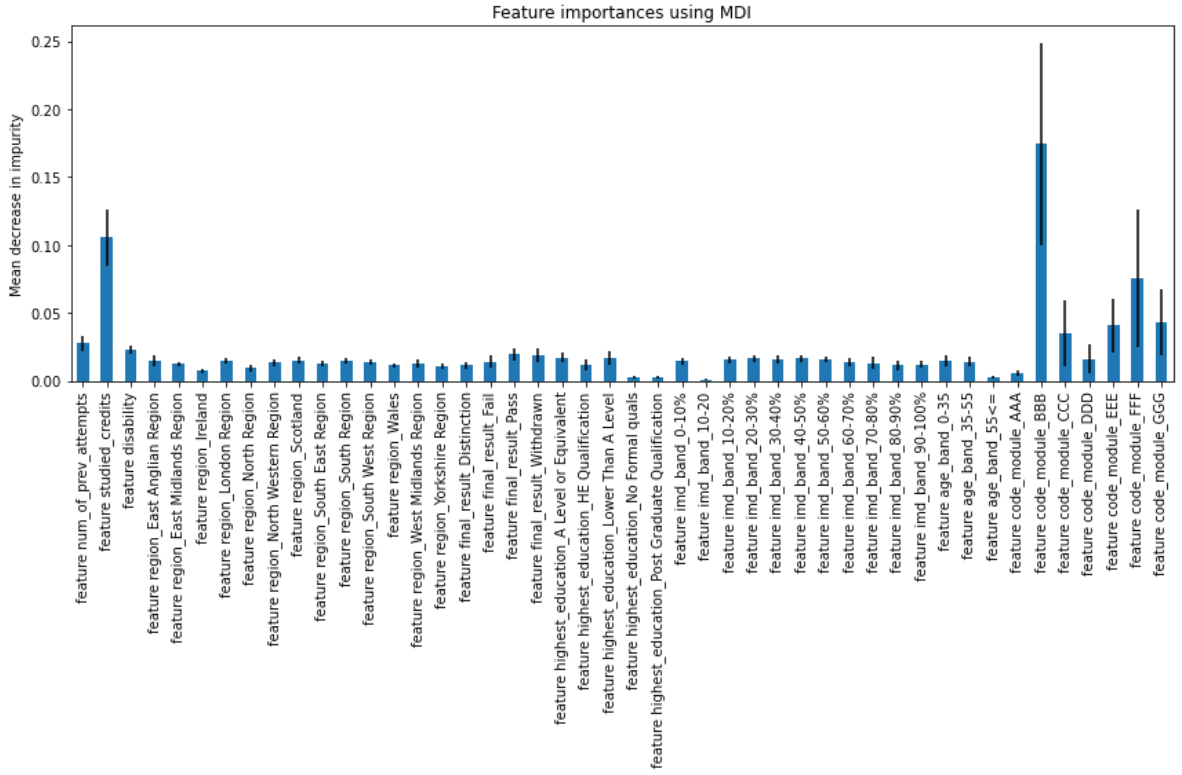

Figure 7: Feature importance showing how classes are the highest predictors along with the number of credits.

When closely looked at, some classes have an extremely uneven distribution when looked at closely by gender. For example, in the highest predictor, module BBB more than 50% of enrolled students are women. Take module FFF, where the opposite happens.

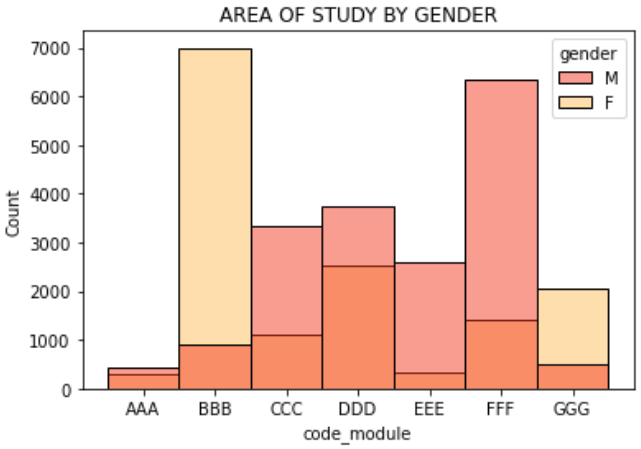

Figure 8: Area of study by gender

Unfortunately, the dataset by Open University does not have a further description for the labels attached to the areas of study and therefore little can be said about specific career

choices and gender. However, what the dataset does show is that there is indeed a clear relationship between gender and certain areas of study at this particular university.

As for the metrics, recall and precision, they are described in the table below:

Class	Precision	Recall
Female	0.77	0.72
Male	0.73	0.79

Table 1: Precision and recall values

Both metrics are around the 70% which shows that the model is both able to identify the category with some degree of accuracy.

5. Discussion & conclusion

From the first question it can be said that Spain follows the trend in career expectation identified by the OECD where girls tend to show less interest in engineering and STEM areas. However, the year 2021 seems to reflect a difference in this dynamic, where the number of students enrolled drop in every area of knowledge, generating women to outnumber men in engineering for example. The change in career choice corresponds to the year following the COVID 19 outbreak, and therefore further study on the causes that affected the trend should be assessed.

In relation to the second question, the feature importance from the model did show there is a strong relationship between gender and areas of study. Nevertheless, gender is a variable that is not easy to predict, and trying to do so based only on demographic and academic information might lead to the trap of perpetuating stereotypes. For this exploratory research, the model was used with the objective of verifying if gender and area of study were strongly related, as a test to the results from the OECD Dream jobs report.

In the case of the open university dataset, the relation between subject choice and gender up being similar to the report in the sense that it showed a strong relationship.

On a side note, finding information related to career choice was demanding, since institutions do not tend to publish information about their students or the enrollment process, and when information is published it is found as a statistical summary. The limitations in terms of data access are detrimental to the understanding of such gender dynamics when it comes to career choice.

References:

1. Michela Carlana, Implicit Stereotypes: Evidence from Teachers' Gender Bias, *The Quarterly Journal of Economics*, Volume 134, Issue 3, August 2019, Pages 1163–1224, https://doi.org/10.1093/qje/qjz008

2. Konrad AM, Ritchie JE Jr, Lieb P, Corrigall E. Sex differences and similarities in job attribute preferences: a meta-analysis. Psychol Bull. 2000 Jul;126(4):593-641. doi: 10.1037/0033-2909.126.4.593. PMID: 10900998.

3. Diekman, A. B., Brown, E. R., Johnston, A. M., & Clark, E. K. (2010). Seeking Congruity Between Goals and Roles: A New Look at Why Women Opt Out of Science, Technology, Engineering, and Mathematics Careers. Psychological Science, 21(8), 1051–1057. https://doi.org/10.1177/0956797610377342

4. Michela Carlana, Implicit Stereotypes: Evidence from Teachers' Gender Bias, *The Quarterly Journal of Economics*, Volume 134, Issue 3, August 2019, Pages 1163–1224, https://doi.org/10.1093/qje/qjz008