National College of Ireland

BSc (Hons) in Computing

Data Analytics

2020/2021

Zara O'Brien

x17363043

x17363043@student.ncirl.ie

Analysis of women in STEM

Technical Report

Table of Contents

Table of Figures

Table of Tables

Executive Summary

Women make up only 27% of the STEM (Science, Technology, Engineering, and Math) workforce as of 2019, despite the fact that employment in STEM has expanded by 79% since 1990. Understanding the variables that contribute to the lack of women in STEM fields is essential if more women are to be inspired to seek careers in the sector.

This study details the CRISP-DM methodology's knowledge acquisition approach. It outlines how the information was gathered and processed for analysis. The investigation aimed to glean information from a STEM-based survey, which questioned STEM professionals regarding the field's vulnerabilities. It detected trends and patterns in students who chose to study a STEM topic in their final year of secondary school by using a SEM (structural equation model).

The study also attempted to predict whether or not a student will choose a STEM topic in school. Many reasons, including a lack of support, role models, interest, and self-efficacy, have been identified as contributing to the absence of women in STEM. The article also includes recommendations for how to address these issues. The data was subjected to a variety of machine learning algorithms, with Multiple Logistic Regression proving to be the most effective.

1.0 Introduction

1.1. Background

STEM stands for Science, Technology, Engineering, and Mathematics, and refers to any topic that can fall under these subjects. The term originates from talks in the United States regarding the shortage of skilled graduates for high-tech jobs. Judith A. Ramaley, who was the associate director for Education and Human Resources at the National Science Foundation from 2001 to 2004, is credited with coming up with the acronym which was originally SMET (Science, Mathematics, Engineering and Technology). To overcome this shortage, governments and institutions around the world have made enticing students to STEM courses a primary focus since its creation. This started in 2000 when the then governor of Texas George W. Bush proposed spending $345 million to increase federal student-loan forgiveness for students who major in Science, Math, Technology or Engineering and commit to teaching in a high-need school for at least five years (Loewus, 2015). The need to increase student uptake in STEM has been a pressing issue for years and is only getting worse. In the foreseeable future, millions of STEM jobs are expected to be vacant. In fact, by 2025, it is expected that 3.5 million jobs would need to be filled in the US (Lazio and Ford, 2019).

STEM courses have traditionally been male dominated, with young girls being deterred from following a technical professional path. Universities, corporations, and governments have all implemented policies targeted at raising the number of women who choose to study in these fields. Only 3% of women in bachelor's degree courses go on to work in the STEM industry, although 12% of women will graduate with a STEM degree each year (O'Callaghan, 2021). To counteract this, government and non-profit campaigns have launched various initiatives to raise the number of young women pursuing STEM degrees. Girls Who Code, STEM Like a Girl and Engineer Girl among several other organizations have sprung up to encourage young women to pursue careers in these fields, with internships and other employment opportunities often made available exclusively for them.

1.2. Motivation

The idea for my project came as a result of my personal experiences. I first gained an interest in technology in my transition year during secondary school when Dell came in to speak to us about careers in IT. After their presentation, they asked anyone who had an interest in a future career in IT to raise their hand and to my surprise, I was the only girl with my hand up. I had thought that college would be different but on my first day, I was surprised at the ratio of women to men. I think there were only about five women in my course in the first year. Last year I was able to do a six-month work placement at Arthur Cox, during my time here I continued to notice this trend with only three out of the fourteen people in the IT team being women. As I'm in my final year I am thinking about work after college and wondering if this trend will persist? Why is it like this? and how to improve it?

Although personally I only have experience with tech industry from some quick research it became very obvious that there is a significant gender gap in all STEM subjects. This is a

topic of much discussion as companies want to try to support minorities to encourage them to pursue careers in areas where they may be outnumbered. Many factors can contribute to the choice to pursue STEM. According to one study, women are 1.5 times more likely than men to drop out of the STEM pipeline after calculus due to a lack of mathematical confidence (Ellis, Fosdick and Rasmussen, 2016). In general, self-efficacy seems to be a big issue, Microsoft discovered that boosting the number of STEM mentors and role models, especially parents, can aid young girls to gain confidence in their ability to thrive in STEM fields, as girls who are supported by their parents are twice as likely to continue in said fields (Choney, 2018). Self-efficacy was a recurring theme throughout my research, in one particular study, it was discovered that, regardless of actual competence or grade in the topic, women consistently rated themselves lower on academic ability in STEM disciplines than their male colleagues (Correll, 2001). These are all things I will try to examine within my investigation.

Before commencing my project, I looked for any similar work to see what the best approach to would be to take in my analysis. Surprisingly there were very few research papers that used data to look at why there is a gender gap in STEM, but I did discover one that I found extremely interesting. Student Factors Influencing STEM Subject Choice in Year 12: a Structural Equation Model Using PISA/LSAY Data (Jeffries, Curtis and Conner, 2020) looked to examine what factors affect a student's decision to do a STEM subject in what would be equivalent to the Irish leaving certificate year. Straight away I realised this paper would be similar to what I was interested in, and I requested a copy of the full paper from the author who was more than happy to send me a copy. When I looked at this research, I realised that I could expand on this by adding in new data and techniques, I also wanted to look at steps that can be taken to try and increase the overall uptake of STEM by women. This method used a SEM (Structural Equation Modelling) model which is typically used to investigate and assess multivariable causal relationships differently from other modelling approaches as they examine both direct and indirect impacts on predetermined causal linkages. This seemed to work well for that problem therefore I will apply that to my research.

1.3. Aims

Objective 1: The initial goal is to create a data warehouse to collect, prepare, and store the data needed to execute the analysis.

Objective 2: The second goal is to examine trends in existing STEM degrees by exploring and analysing the data I've accumulated.

Objective 3: To learn about existing STEM workers' perspectives on issues in the sector and why they believe there is a gender imbalance.

Objective 4: To develop a SEM model to investigate the links between mediating factors affecting students' choices of STEM subjects for their curriculum.

Objective 5: The next aim is to determine whether a prediction is achievable and to assess several models for making predictions using different predictive modelling algorithms.

Objective 6: After I've gathered all of my results, I will present and visualize them using a variety of graphical tools, then develop an interactive dashboard from them. From the results, we can also identify the most significant barriers that prohibit women from pursuing careers in STEM fields and devise solutions to these issues.

1.4. Technology

The following technologies were used to achieve my results:

R Language: R is a programming language for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. I will be using R for data preparation and to create Logistic Regression, Naïve Bayes, Decision Tree and Random Forest models.

Fig. 1 R Logo

R Studio: I will be using R Studio to build my project. R studio is an open-source integrated development environment (IDE) for R Language. R will be used to prepare the data as well as analyse and apply models to the data. I utilised many libraries in my project including RODBC, ODBC, DBI, Lavaan, semPlot, Caret, Tidyverse, Dlookr, Rsample, Ggplot2, e1071, C50, rpart, rpart.plot, RColorBrewer, ROCR, rattle, gbm, adabag, MASS, randomForest and Hmisc.

Fig. 2 R Studio Logo

Tableau: Tableau is a free data visualization software. I will use tableau to display all of my findings and my final figures. Tableau dashboards are visually appealing and user friendly, therefore I think it is the best option for presenting my results.

Fig. 3 Tableau Logo

SPSS Statistics: SPSS Statistics is a software package used for interactive, or batched, statistical analysis and is one of the most popular statistical tools. I will use SPSS to look at my data and perform various statistical tests to ensure all the applicable data assumptions are not violated. SPSS will also be used to generate basic descriptive statistics.

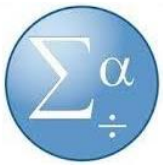


Fig. 4 SPSS Statistics

SPSS Amos: SPSS Amos is a comprehensive structural equation modelling software that extends basic multivariate analytic methods like regression, factor analysis, correlation, and analysis of variance to assist your research. SPSS Amos will be used to construct, visualize, test, and optimise the SEM model for the analysis.

Fig. 5 SPSS Amos Logo

Microsoft SQL Server: Microsoft developed a relational database management system which is a software that serves as a database server for storing and retrieving data as required by other software programs. All my data will be stored here and retrieved by R Studio, Alteryx and Tableau as needed.

Fig. 6 Microsoft SQL Server Logo

Alteryx: Alteryx is a program for manipulating data and reads data directly from SQL Server, as such it will be used in my project for data preparation. Specifically, it will be used to remove all noise from the datasets and for different cleaning challenges. Alteryx will also be used for dealing with missing data by both removal and imputation.

Fig. 7 Alteryx Logo

Microsoft Office Suite: I will be using Excel for my initial look at the data and to look at it during the data exploration stage of my project. I will use PowerPoint for making the slides for my presentation and I will be using Word to write this document.

Fig. 8 Microsoft Office Suite Logo

For my project, I will follow the CRISP-DM (CRoss-Industry Standard Process for Data Mining) the most widely used methodology since 2002. The reason I chose this methodology over KDD (Knowledge Discovery in Databases) or SEMMA (Sample, Explore, Modify, Model and Access) is that transitions between stages in CRISP-DM can be reversed, which is a significant difference from the two other techniques. When working with real data, this is quite useful. KDD and SEMMA are nearly comparable since every KDD stage equates to a SEMMA stage. The CRISP-DM method integrates the steps of Selection-Pre-processing (KDD) and Sample-Explore (SEMMA) into the Data Understanding stage. It also includes the steps of Business Understanding and Deployment which would be useful as I want to try and figure out what steps we can take to encourage more girls to take up STEM. I will describe the six stages of the CRISP-DM methodology below.

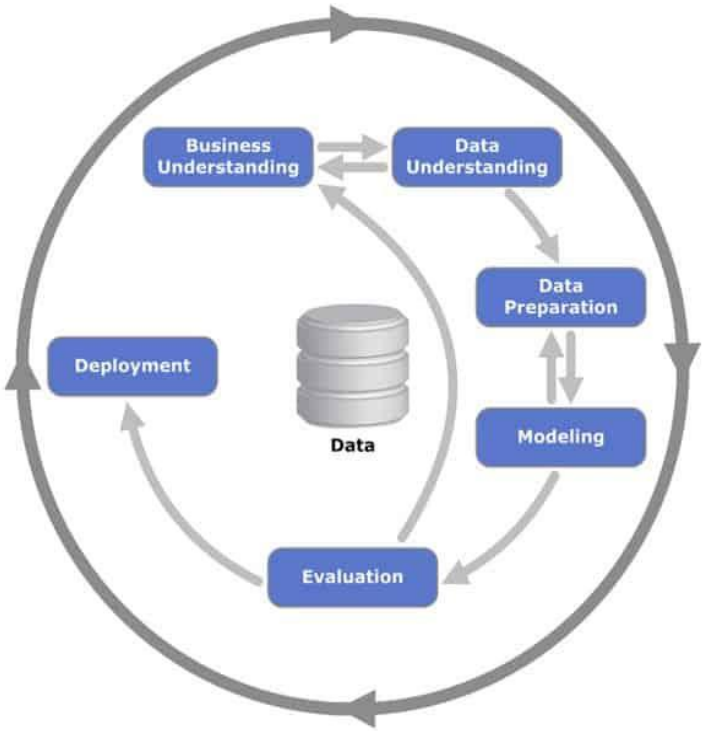


Fig. 9 Stages of CRISP-DM

Business understanding: This involves making a preliminary plan for the project and understanding the aims/objectives of the project. Within these objectives contains information regarding the background of the investigation as well as success criteria, similar to a definition of Done (DoD) which is the minimum number of deliverables required to consider the project completed, part of this is defining my goals which I set out to achieve at different phases of the analysis giving me a tangible measure of progress. I will also take this time to assess my available resources and also to weigh the benefits or problems associated with the project as well as any assumptions made about the investigation. Finally, once all of this information is aggregated, I will create my final project plan.

Data understanding: This step involves the collection of all the data required for the project and looking at the data to get first insights into what the data could show. The first step is to collect the data on which I will conduct the analysis, from there we begin to characterize the data and highlight important aspects and meta-information about the data. Once I am comfortable with the broad concepts of the dataset, I begin to navigate the detailed structure of the data and from there I can examine the quality, feasibility, and appropriateness of the data regarding the business objectives.

Data preparation: This is the process of constructing the final dataset to be used. This will involve the cleaning, preparation, and storage of the data. This step is a very sensitive and important aspect of the overall project, data preparation significantly influences the success of our investigation, and I must take into consideration many different factors as the data is manipulated. I must justify the inclusion and removal of certain data, I must clean it without

altering the contents, I must perform feature engineering to align the data with my objectives, the data must be successfully merged without conflict and correctly formatted.

Modelling: This stage is where any modelling techniques are applied to the dataset. Modelling helps me to conceptualize the data I have prepared to allow me to draw insights and observations. There are a wide variety of modelling concepts and techniques, and the initial task of the modelling phase is to identify the most appropriate tools for my project. Building a model requires a large amount of configuration and development to complete the task it was designed for. Following the creation of a model, its performance is assessed, and the model is optimised if needed to ensure the best possible results.

Evaluation: In the evaluation stage I will check to ensure that my project has properly achieved the objectives and I will review the work to see if anything needs to be changed. By comparing the results against my business success criteria, I can determine whether I can approve my model. If my modelling phase is successful, I can begin to establish the steps going forward.

Deployment: Once I am happy with the results this is where I will actually create the model and display the results or findings of the project. Assuming that I am satisfied with the results of my model, I will present my findings, outline the success of the project, the objectives achieved and summarise the investigation and its results.

1.5. Structure

The remainder of the report is organised as follows:

Section 2: Describes in detail all the data as well as a breakdown of the exact data used in the analysis along with information on where the data was sourced from. Data summary and descriptions will be shown here as well as statistics and all exploratory analysis. This occurs in the data understanding phase of the CRISP-DM framework.

Section 3: Details all pre-processing such as cleaning, transforming, and dealing with missing data and overall, how the data was prepared to be analysed. This processing covers the data preparation stage of the project.

Section 4: Explains all approaches taken in the analysis along with why these were selected and how exactly they were implemented. This section falls within the modelling phase of the CRISP-DM framework.

Section 5: Following the completion of the modelling phase, the evaluation phase commences. In this section, I present the results and evaluation metrics from the analysis along with findings and any tables and figures.

Section 6: Finally, the deployment phase concludes the CRISP-DM framework. Here we discuss the key findings of the project as well as how they can be utilised. The strengths and limitations of the project will also be discussed here.

Section 7: Addresses future work and development on the project with the addition of time and resources.

Section 8: Shows all references in the document.

Section 9: Is the appendix to the report.

2.0 Data

In this research, I use three data sets. The first dataset is a relatively small table from the U.S Department of Education in the National Centre for Education Statistics. The National Center for Education Statistics (NCES) is the major federal agency in charge of gathering and evaluating educational data in the United States and other countries. The table titled 'Number and percentage distribution of science, technology, engineering, and mathematics (STEM) degrees/certificates conferred by postsecondary institutions, by race/ethnicity, level of degree/certificate, and sex of student: 2008-09 through 2017-18' (National Center for Education Statistics, 2019) details the total number of stem degrees given by year and gender which is useful for me to get an idea of the overall trend. This data is available publicly and freely to download and use therefore I did not need to apply for permission to use it.

The second data set used in this report is from the Pew research centre which is "... a nonpartisan fact tank that informs the public about the issues, attitudes and trends shaping the world". This is a STEM survey from 2017 which investigates the trends and differences related to gender in STEM. The data was used to write the report titled 'Women and Men in STEM Often at Odds Over Workplace Equity' (Funk and Parker, 2018). This report focused on issues in the workplace but the data behind it would be useful to gain an insight into the feelings of the people working in STEM on the issues. This data is also publicly available online with the disclaimer that "Pew Research Center bears no responsibility for the analyses or interpretations of the data presented here. The opinions expressed herein, including any implications for policy, are those of the author and not of Pew Research Center" (Pew Research Center, 2017).

The last and perhaps the most important data set used in this analysis is the LSAY (Longitudinal Surveys of Australian Youth) dataset. This dataset gathers data about training and education, work, financial matters, health, social activities, and other issues from broad, nationally representative children at school. The survey has been linked to the PISA (Programme for International Student Assessment) dataset since 2003. Data is primarily gathered using a mixture of student academic assessments and a school-based survey. Annual telephone interviews are used to collect further data. Survey participants have had the opportunity to finish their interviews online since 2012. This data allows us to see if students chose a STEM subject along with lots of other information. Access to the LSAY data is free via a formal request and registration process managed by the Australian Data Archive (ADA), I applied for data and gained access shortly after. The data sets run over 10 years, I

used the latest completed data set which covered 2009 – 2019 (Australian Government Department Of Education, 2020).

I initially explored all data by opening it up in either SPSS or Excel and seeing how clean and complete it was. I also got an idea of what data I would not need and generated some descriptive statistics to get an insight into the state of the data. The table below shows a summary of the data when downloaded in its original state.

Data File	File Format	File Size (KB)	Number of Records	Number of Attributes	Attribute types
Degrees	Excel (.xls)	18,811	180	20	Numeric
Pew	SPSS (.sav)	1,215	4,914	220	Numeric
LSAY	SPSS (.sav)	131,069	14,251	7265	Numeric

Table 1 Data description

As you can see in Table 1 there are a lot of attributes in the data, particularly in the Pew and the LSAY data, not all of these will be useful or of any interest in this research so should be removed accordingly. This was done at this stage as there is a column size limit of 1,024 which the LSAY data exceeded. I wanted to upload all my data to the Microsoft SQL server before starting any cleaning or preparation on it but due to the restrictions on size, noise was removed from data before loading it into the SQL Server. The noise was removed from the LSAY data using Alteryx and was then loaded into the SQL server along with the other two data sets. At this stage, I decided to create my own excel sheet to keep track of what data remained for the analysis along with their definitions and coding details. Below you can see the breakdown of the data that was kept for the analysis.

Column Name	Description
Year	Year conferred
TotalSTEM	Total number of Bachelor STEM degrees
MSTEM	Total number of Bachelor STEM degrees given to Males
FSTEM	Total number of Bachelor STEM degrees given to Females

Table 2 Degree data description summary

Column Name	Description
CaseID	Unique respondent identification number
PPGENDER	Gender
SCH7	What's the main reason many young people don't pursue college degrees in STEM?
WORKTYPE_FINAL	STEM Status
SCH10A_1	I found science classes easy - Y/N
SCH10A_2	It was easy to see how science would be useful for the future – Y/N
SCH10A_3	I felt that I belonged in science classes – Y/N
SCH10A_5	I had a lot of support at home or after school to help me do well in these classes – Y/N
SCH10B_1	I found science classes hard – Y/N
SCH10B_2	It was not easy to see how science would be useful for the future – Y/N
SCH10B_3	I felt that I didn’t belong in science classes – Y/N
SCH10B_5	I didn’t have enough support at home or after school to help me do well in these classes – Y/N
TALENT	For the kind of work that you do, how important, if at all, would you say having a natural ability is or has been for you, personally, to get ahead in your job?
FAMSTEM1	Do you have any close family members who work or have worked in a job or career that involves STEM?

FAMSTEM2_1	These close family members who work or have worked in a job or career that involves STRM are they older or younger?
INTEREST1	Were you, personally, ever interested in pursuing a job or career that STEM?
REASON1a	From an early age, girls are not encouraged to pursue these subjects in school?
REASON1b	Women are less likely than men to believe that they can succeed in these fields?
REASON1c	Women do not pursue these jobs because there are so few female role models in these fields?
REASON1f	Women are just less interested in STEM than men?
GENDDISC_f	Have you had someone treat you as if you were not competent because of your gender?

Table 3 Pew data description summary

Column Name	Description
STIDSTD	Student ID
SECTOR	School sector
ST04Q01	Sex
IMMIG	Immigration status
ST62N01	How well doing - English
ST62N02	How well doing - Maths
ST62N03	How well doing - Science
ST62N04	How well doing - Subjects overall
ST67N01	Views on sci - study at sec sch

ST67N02	Views on sci - study after sec sch
ST67N03	Views on sci - work in career
HISCED	Highest educational level of parents
ATTCOMP	Attitude towards computers
ATSCHL	Attitude towards school
HIGHCONF	Self-confidence in ICT high level tasks
WEALTH	Wealth
PV1MATH	Plausible value in math
PV2MATH	Plausible value in math
PV3MATH	Plausible value in math
PV4MATH	Plausible value in math
PV5MATH	Plausible value in math
PV1SCIE	Plausible value in science
PV2SCIE	Plausible value in science
PV3SCIE	Plausible value in science
PV4SCIE	Plausible value in science
PV5SCIE	Plausible value in science

Table 4 LSAY data description summary

As you can see in Table 2, Table 3 and Table 4 the data was filtered down to what I believed to be useful for the study along with any data that was recommended in the paper which used older LSAY data (Jeffries, Curtis and Conner, 2020). The LSAY data has nearly 200 additional columns that will be used for feature engineering but are then removed from the data set so are not included in Table 4.

3.0 Methodology

3.1. Exploratory Analysis

Before preparing the data, I wanted to do an exploratory analysis of both the Degree and Pew dataset to get insights into what could be useful to look at in the LSAY dataset. With the Degree dataset, all I could do was plot the data to see if there were any trends. As seen in the below graph Fig. 10 crated in Tableau it can be seen that although the number of bachelor's degrees being achieved each year is growing the gender gap within these is still prevailing. This shows that this is likely to remain an issue and is something that should be examined in attempt to improve it.

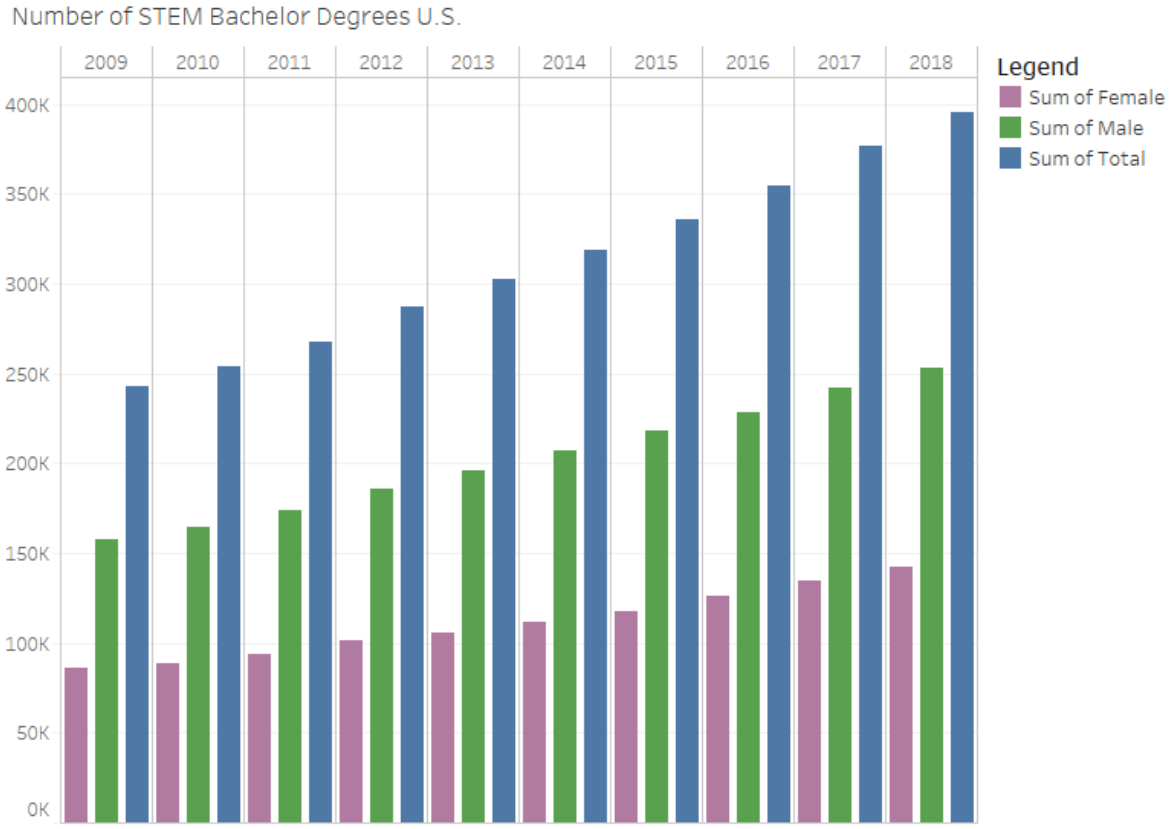


Fig. 10 Bar chat of the number of STEM degrees in the U.S by year and sex

After this I started to look at the Pew dataset in R. First the data set was loaded in from SQL Server. Then after doing some quick descriptive statistics the data was first split into two new data frames by each person's gender and this was filtered further to people who work in STEM the code to complete this can be seen in Fig. 11. Each group was saved into a new data frame allowing answers to be compared between sex and STEM status. Once this was done, each question of interest was individually examined using the describe function in R or working out the percentage of people who answered a certain way. Any insights that were drawn from the data helped guide me with what to look for in the LSAY data. Some of these insighted are listed below.

```
# Split data by gender ----------------------------------------
pewM <- pew[pew$PPGENDER == 1,] # 1 = Male
pewF <- pew[pew$PPGENDER == 2,] # 2 = Female


# Filter to people working in STEM ----------------------------
pewMSTEM <- pewM[pewM$WORKTYPE_FINAL == 1,] # 1 = STEM Worker
pewFSTEM <- pewF[pewF$WORKTYPE_FINAL == 1,] # 1 = STEM Worker
```

Fig. 11 Filtering Pew data

- 55% of all people surveyed believed that the main reason many young people don't pursue college degrees is that they think these subjects are too hard.
- 43% of men working in STEM found their primary school science classes easy compared to 38% of women.
- 6% more men could see how science would be useful for the future than women.
- 47% of men felt they belonged in science classes and 39% of women did.
- 23% of both men and women felt they had a lot of support at home.
- 43% of both sex's thought that girls not being encouraged to pursue these subjects in school from an early age is a major reason for the lack of women in STEM.
- When asked have you had someone treat you as if you were not competent because of your gender, 3% of men working in STEM said yes compared to 29% of women.

3.2. Pre-processing

Overall, the Degree data and the Pew data sets did not need pre-processing or preparation as they were mostly used in the exploratory analysis stage of the project, however, the LSAY required quite a lot of preparation before it could be used in the analysis including checking for certain assumptions needed for the modelling stage of the report. As stated above any noise was removed from the LSAY data set before being uploaded to the SQL server, this was completed in Alteryx as seen in Fig. 12 using their built in select function and deselecting any variables we do not wish to keep before saving it in a CSV file.

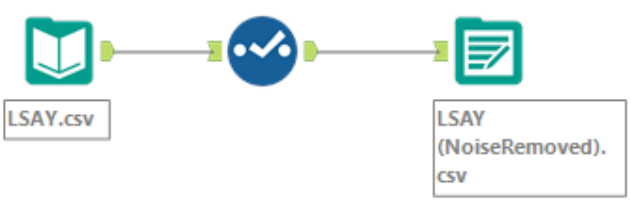


Fig. 12 Alteryx remove noise workflow

As mentioned previously in the report all data was stored in a SQL Server as seen in Fig. 13. The first step in preparing the data is to load it in from the server using the RODBC package, then I created a copy of the dataset just to make it easier to roll back should things not go to plan.

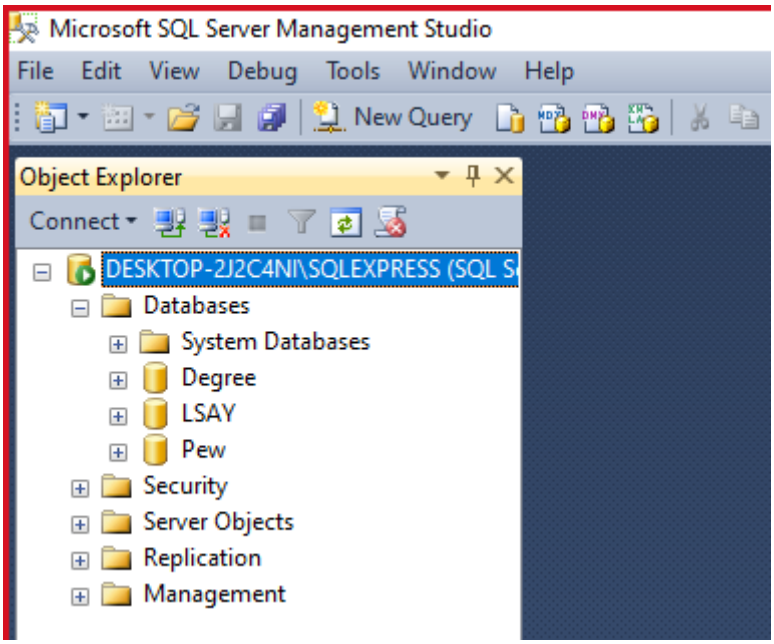


Fig. 13 SQL Server databases

The first thing that needed to be done to the data was feature engineering, one of the aims of this study is to apply predictive models to the data to see if forecasts can be made. For this, the outcome or response variable is if a student took a STEM class in their final years of school which would be equivalent to the Irish leaving certificate. This information is not directly given in the dataset but can be created using a large combination of columns. The data contains a new column for every subject available asking if the student took that subject. To know what subjects to include as STEM I used the STEM Designated Degree Programs list (DHS, 2016). This is a comprehensive list of disciplines of study that the Department of Homeland Security considers being science, technology, engineering, or mathematics (STEM) fields of study.

```
# Create STEM Var
LSAYCopy$STEM <- c(0)

# ISSTEM - In School STEM
ISSTEM = data.frame()

# Function for adding students who said yes to doing a STEM subject to ISSTEM
Schoolfunction <- function(col){
  for(i in 1:nrow(LSAYCopy)) {
    if(!is.na(col[i])){
      if(!is.na(match(col[i], 1))){
        ISSTEM <<- rbind(ISSTEM, LSAYCopy[i,])
        print(col[i])
      }
    }
  }
}
```

Fig. 14 Function to create new STEM variable

In Fig. 14 above you can see the process of creating a new variable called 'STEM' which indicates if a student is doing at least one STEM subject with a 0 denoting they do not and a 1 meaning that they do. Firstly, a new column is added to the data set named STEM and all values are assigned to 0. Next, an empty data frame is created called ISSTEM, here is where all the students who have picked a STEM subject will be added. A function is then created which iterates through each row in the dataset checking that the column is not empty and then checking if the value in the column is equal to 1 using the match function. If this is true, this means that the student is doing that subject and the row is added to the ISSTEM data frame using the rbind function. The print statement was used for testing the function. This function was then applied to all applicable columns.

```
# Remove Duplicates
ISSTEM <- ISSTEM[!duplicated(ISSTEM), ]

# List of students ID who do STEM from ISSTEM
STEMID <- list()
for(i in 1:nrow(ISSTEM)){
  STEMID[i] <- ISSTEM$STIDSTD[i]
  print(ISSTEM$STIDSTD[i])
}

# Loop to assign 1 to LSAYCopy$STEM for all obs in ISSTEM
counter <- 0
for(i in 1:nrow(LSAYCopy)){
  if(!is.na(match(LSAYCopy$STIDSTD[i], STEMID)))
  {
    counter <- counter + 1
    print(counter)
    LSAYCopy$STEM[i] <- c(1)
  }
}
```

Fig. 15 Code to remove duplicates and assign 1 for STEM students

Once the function runs all columns necessary it is important to remove duplicates as many students may do more than one STEM subject e.g., math and science, this is done using the duplicated function. When all duplicates are removed, a list is created called STEMID which contains all of the IDs of students that do a STEM subject. This list is then used in a loop which runs through all rows in the dataset, checking if they are not empty and then if student ID matches the ID's saved in the STEMID list a 1 is assigned to the STEM column for that student's row in the dataset. The counter reached 7,686 showing how many students took a STEM subject in the data set.

```
# List of STEM codes taken from the Australian governments STEM fields of education and research
ANZSCOSTEM <- list(1349, 1311, 1342, 1330, 1349, 1213, 2421, 2356, 2120, 2434, 3151, 3154, 2161,
2162, 2165, 2166, 2145, 2149, 2142, 2149, 2151, 2152, 2164, 2141, 2144, 2146, 2143, 2132, 2113, 2145
, 2133, 2114, 2131, 3212, 2250, 2621, 2146, 2112, 2111, 2269, 2146, 2269, 2265, 3211, 2263, 3257,
2267, 2262, 3253, 2263, 3214, 2269, 2230, 2261, 2269, 2264, 2266, 2211, 2212, 2222, 2221, 2511, 2513
, 2512, 2514, 2519, 2521, 2529, 2522, 2523, 2519, 2153, 2634, 2632, 3142, 3143, 3259, 3211, 3259,
3212, 3259, 3213, 5329, 3214, 3256, 3359, 3257, 3351, 3257, 3359, 3111, 3141, 3143, 3240, 3111, 3119
, 3118, 3112, 3123, 3113, 3155, 2149, 3117, 3257, 3116, 7422, 3512, 3513, 3514, 3511, 3522, 7412,
7231, 8122, 7212, 7211, 7224, 7213, 7212, 7214, 7411, 7412, 7413, 7421, 7422, 3240, 5164, 3254, 3258
, 3251, 3214, 3221, 3222, 3251, 5329, 3255, 3259, 4132, 3252, 3135, 7213, 7223, 7224, 8121, 8122,
1332, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2610
, 2613, 2633, 2241,2312, 3124, 3125, 3129, 3210, 3200, 3220, 3223, 3230, 3231, 3232, 3233, 3234,
3241, 3242, 3243, 3400, 3411, 3422, 3421, 3923, 3933, 3941, 3991, 3992, 1351, 2300, 2310, 2311, 2322
, 2349, 2600, 2611, 2612, 2630, 2631, 3100, 3110, 3114, 3120, 3122, 3130, 3131, 3132, 3423, 3424,
3620, 2347, 2515, 2524, 2525, 2526, 2527, 2530, 2531, 2532, 2533, 2534, 2535, 2539, 2540, 2541, 2542
, 2543, 2544, 3613, 4112, 4114, 4116, 2500, 4111, 4232)
```

Fig. 16 List of STEM ANZSCO codes

The same process was used to create a ParentSTEM variable to see if at least one of the student's parents works in STEM, similarly, this is coded as 0 for no and 1 for yes. To create this variable the same process was followed as explained above. The only difference is that instead of checking if they said yes to doing STEM subjects we checked to see if their ANZSCO code matches a list of STEM ANZSCO codes as seen in Fig. 16. This list was created using the Australian Government's given STEM fields of education and research (Australian Government, 2021). ANZCSO codes are the Australian and New Zealand Standard Classification of Occupations.

```
for(i in 1:nrow(LSAYCopy)){
  # Add 0 to numbers so there will be no duplicates when changing.
  if(!is.na(match(LSAYCopy$ST62N01[i], 1)))
  {
    LSAYCopy$ST62N01[i] <- c(10)
  }
```

Fig. 17 Adding 0 to the end of numbers

One thing I noticed when exploring the data was that not all variables were of the same scale, meaning that in most cases a higher number indicated a more positive response e.g., 1 = bad, 2 = average and 3 = good. However, there were a few variables that did not follow this format and were in the reverse. It was decided it would make the most sense to change

the scale of these variables so that they were all the same. The first step in doing this was to add a 0 to the end of all numbers in the applicable column as seen in Fig. 17, this was to avoid confusion when changing scale. If this step was not taken after switching the first number for example changing all the 3's to 1's there would have been two sets of 1's that could not be differentiated. The scale was then changed using a loop running through each row for all applicable columns as seen in Fig. 18.

```
for(i in 1:nrow(LSAYCopy)){
  # Change Scale
  if(!is.na(match(LSAYCopy$ST62N01[i], 10)))
  {
    LSAYCopy$ST62N01[i] <- c(5)
  }
```

Fig. 18 Change scale of variables

One thing that needed to be dealt with was how the data was coded, for all variables 7 = NA, 8 = Invalid and 9 = Missing. For the sake of the analysis, I counted how many of each were contained in each variable and found that there were only 9's which is missing data. As previously I had changed the scale so that a higher value would denote a more positive response it would not make sense to keep these coded like this, so they were all recoded to NAs for the seven variables that contained these responses. As seen in Fig. 19 a loop was created to check these columns for the value of 9 and if found would change it to NA, a counter was also added along with a boolean flag to count how many unique student ID's would be changed.

```
counter = 0

  for(i in 1:nrow(LSAYCopy)) {

    found <- FALSE

    if(!is.na(match(LSAYCopy$ST62N01[i], 9))){
      LSAYCopy$ST62N01[i] <- c(NA)
      found = TRUE
    }
```

Fig. 19 Changing values equal to 9 to NA's

3.3. Assumption Checks

At this point, all variables that were used for the feature engineering were removed as they would not be used again, and the data was saved for checking assumptions of the data. As my starting goal of the project was to create a SEM model, I checked for the assumptions of this. The major assumptions associated with SEM include multivariate normality, no systematic missing data and sufficiently large sample size (Bentler, 2001). To keep track of

everything I created an excel sheet for all variable information as seen in Fig. 20 this way I could keep track of all test results and information about each variable.

An alpha value of 0.05 was declared for the purpose of testing assumptions. This means I am willing to accept a 5% chance of making a type 1 error, which would result in rejecting the H0 (null hypothesis) when it should be kept. This alpha was chosen since it is the most generally used value, and I did not see the need to extend it to 0.01 because this is solely for checking assumptions.

	A	B	C	D	E	F	G	H	I	J	K
1	**Column**	**Description**	**Values**	**Mean**	**Mode**	**Median**	**Skewness**	**Kurtoisis**	**Miss val %**	**VIF**	**Outliers**
2	SECTOR	School sector	1 Catholic, 2 Go	1.95	2	2	0.046	-0.556	0	1.02	
3	ST04Q01	Sex	1 Female, 2 Male	1.48	1	1	0.097	-1.991	0	1.063	
4	IMMIG	Immigration stat	1 Native, 2 Seco	1.29	1	1	1.965	2.377	2.1	1.053	
5	ST34Q01	Teachers - Get a	1 Strongly disagr	3.03	3	3	-0.63	1.351	2.2	1.897	
6	ST34Q02	Teachers - Inter	1 Strongly disagr	2.88	3	3	-0.637	1.081	2.6	2.172	
7	ST34Q03	Teachers - Real	1 Strongly disagr	2.79	3	3	-0.52	0.483	2.5	2.313	
8	ST34Q04	Teachers - Extra	1 Strongly disagr	3	3	3	-0.621	1.474	2.5	1.822	
9	ST34Q05	Teachers - Trea	1 Strongly disagr	2.99	3	3	-0.704	1.786	2.4	2.079	
10	ST62N01	How well doing -	1 Very poorly, 2 M	23.03	20	20	-0.027	-1.123	0	6.531	
11	ST62N02	How well doing -	1 Very poorly, 2 M	24.86	30	30	-0.365	-0.966	0	10.554	
12	ST62N03	How well doing -	1 Very poorly, 2 M	25.65	30	30	-0.441	-0.932	0	7.637	
13	ST62N04	How well doing -	1 Very poorly, 2 M	3.82	3	4	2.056	6.738	0	1.194	
14	ST67N01	Views on sci - st	1 Strongly disagr	3.22	2	3	1.659	2.016	0	10.277	
15	ST67N02	Views on sci - st	1 Strongly disagr	3.39	3	3	1.643	2.032	0	15.397	
16	ST67N03	Views on sci - w	1 Strongly disagr	3.51	3	3	1.624	2.025	0	11.559	
17	HISCED	Highest educati	1 ISCED 1 - Prim	4.79	6	5	-0.464	-0.673	4.2	1.17	
18	ATTCOMP	Attitude towards	9997 N/A, 9998	-0.158683	0	-0.108	-0.601	-0.051	6.5	1.085	
19	ATSCHL	Attitude towards	9997 N/A, 9998	0.173841	-0.5	0.0185	0.269	-0.473	5.6	1.295	13819, 13049, 1384
20	HIGHCONF	Self-confidence	9997 N/A, 9998	0.129934	2	-0.0459	0.003	0.298	5.9	1.101	11725, 11715, 11886
21	WEALTH	Wealth	9997 N/A, 9998	0.726986	1	0.7032	0.589	1.025	1.3	1.064	13568, 5559, 5405
22	PV1MATH	Plausible value i	NA	519.2722	500	520.26	-0.068	-0.95	0	9.507	14171, 5422, 658, 4
23	PV2MATH	Plausible value i	NA	518.6361	550	519.4	-0.087	-0.14	0	9.719	2391, 2551, 3033,
24	PV3MATH	Plausible value i	NA	519.0984	550	520.88	-0.074	-0.08	0	9.304	2551, 14163, 090, 1
25	PV4MATH	Plausible value i	NA	518.5581	550	520.1	-0.064	-0.104	0	9.976	8206, 5422, 1447,
26	PV5MATH	Plausible value i	NA	519.4104	560	520.88	-0.072	-0.098	0	9.67	14142, 1915, 14163,
27	PV1SCIE	Plausible value i	NA	533.6244	560	535.94	-0.128	-0.028	0	11.656	1882, 14171, 14159,
28	PV2SCIE	Plausible value i	NA	533.1822	550	536.22	-0.134	-0.021	0	11.727	14152, 7315, 13873
29	PV3SCIE	Plausible value i	NA	533.6007	570	535.94	-0.139	-0.02	0	11.56	14164, 14152, 6214,
30	PV4SCIE	Plausible value i	NA	533.3298	570	536.22	-0.127	-0.011	0	11.745	14142, 11404, 1413
31	PV5SCIE	Plausible value i	NA	533.5938	510	536.68	-0.122	-0.062	0	12.001	14142, 14153, 1415
32	STEM	People who cho	0 No, 1 Yes	0.59	1	1	-0.383	-1.854	0	1.26	
33	ParentSTEM	People who hav	0 No, 1 Yes	0.23	0	0	1.314	-0.273	0	1.065	

Fig. 20 Variable information excel sheet

The first tests were for data normality. The hypothesis for this test are:

H0: The data is normal

H1: The data is not normal

Data normality is a rare occurrence in real data. Skewness and kurtosis functioned as indicators of normality for this study. The Kolmogorov-Smirnov test for normality is likely to detect non-normality because the sample size is big (> 200) similarly to the Shapiro-Wilk as the sample size is (>5,000). Descriptive statistics can be used to investigate both skewness and kurtosis, these were generated in SPSS and saved to be examined further Fig. 20 . When using SEM, acceptable skewness values are between 3 and + 3, and acceptable kurtosis values are between 10 and + 10 (Brown, 2006). The descriptive statistics show that the skewness of the data is between (-0.704 and 2.056) and the kurtosis is between (-1.991 and

6.738) meaning we can accept H0 and reject H1 as It would appear that the data was statistically proven to be normal.

The data was then tested for multicollinearity. The hypothesis for this test are:

H0: VIF < 5

H1: VIF > 5

Using linear regression and collinearity diagnostics in SPSS, multicollinearity between independent variables was investigated. The presence of multicollinearity among manifest variables can be problematic. Values of the variance inflation factor (VIF) greater than 5 should be investigated further and values greater than 10 should be excluded from the analysis. All VIF values greater than or equal to 5 or 10 will be used to generate a latent variable, that should solve the multicollinearity problem; all other variables are within the allowed range therefore we can accept H0 and reject H1 in the case of observed variables that will not be used to create new latent variables.

The linearity assumption was tested next, using the following hypothesis:

H0: Data is linearly correlated

H1: Data is not linearly correlated

Some statistical techniques, such as SEM, are based on the assumption that the variables are linearly correlated. As a result, visualizing the data in a scatterplot is a common method to see the coordinate pairs of data points. The linearity assumption was tested using correlation coefficients and scatterplots between pairs of scale variables. These plots were made in SPSS and stored for further evaluation. Upon examining the scatterplots, I could accept the H0 and reject the H1 as I felt this assumption was satisfied.

When testing for outliers the following hypothesis were evaluated:

H0: There are no outliers

H1: There is at least one outlier

The data was next checked for univariate and multivariate outliers, both of which might have an impact on the results of statistical studies. Outliers can have a significant impact on structural models in a variety of contexts, including SEM and regression. Outliers can be caused by observational errors, data entry problems, or genuine extreme results from self-reported data. Outliers must be explained, removed, or accommodated since they alter the mean, standard deviation, and correlation coefficient values. A data point that consists of an extreme value on one variable is referred to as a univariate outlier. A multivariate outlier is defined as a set of unusual results over at minimum two variables. Because scale values

have so-called floor and ceilings, they cannot have outliers, but continuous values can, and they should be deleted. To look for multivariate outliers, SPSS was used to create a new column that contained the Mahalanobis Distance for each variable, which was then sorted in descending order with the larger values appearing first. Then the Mahalanobis Distances are compared to a chi-square distribution with the same number of degrees of freedom. The p-value of the test will be placed in another new column named Probability_MAH_1. Wherever the new probability variable's values are less than .001, multivariate outliers would appear; this indicated 290 multivariate outliers that need to be removed. Boxplots in SPSS were used to look at univariate outliers, revealing 96 more outliers. I made a list of student numbers which were outliers, both multivariate and univariate, and eliminated them in R studio. We can tell by the above tests there was a lot of outliers in the data set and therefore we must reject H0 in favour of H1. Once the outliers were removed the data was saved back to SQL Server using the code in Fig. 21.

```
# connect to a local MS SQL Express instance
con <- DBI::dbConnect(odbc::odbc(),
                      Driver = "SQL Server",
                      Server = "localhost\\SQLEXPRESS",
                      Database = "LSAY",
                      Trusted_Connection = "True")

# Overwrite data
dbWriteTable(con, "LSAY", LSAYCopy, overwrite = TRUE)
```

Fig. 21 Write processed data back to SQL Server

Because SEM may describe complex relationships between multivariate data, the sample size is an important but often overlooked consideration. Two commonly held beliefs are that you'll need more than 200 observations or at least 50 times the number of constructs in the study. For SEM, higher sample size is always preferred. With n variables, the formula is u = 1/2 n (n+1). In the case of non-normal data, u denotes the elements required to form a model. Larger samples are required when data is not regularly distributed or is otherwise problematic in some way. When data is skewed, kurtotic, sparse, or otherwise less than optimal, it's tough to make clear recommendations about sample sizes. As a result, the conventional advice is to collect additional data wherever possible. After removing all outliers from 14,251 observations, 13,865 remain, indicating that the data size criterion has been reached.

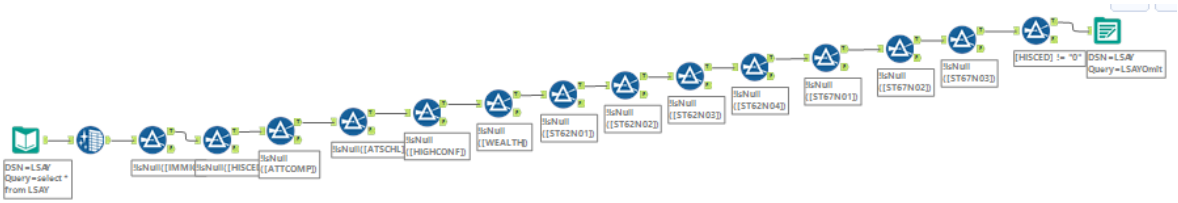


Fig. 22 Alteryx workflow for listwise deletion

When evaluating missing data, we must first determine whether any variables are missing at random (MAR), missing completely at random (MCAR), or not missing at random (NMAR). Variables in the study should have all of their data forms filled out or no data should be missing in any variable. Listwise deletion of cases, where an entire case's record is erased if the case contains one or more missing data points, is a common ad hoc solution to missing data problems. The substitution of the variable's mean for the missing data points on that variable is another commonly used ad hoc missing data management strategy. I decided to do both, keeping one copy of the data with only full cases and another with missing data imputed using the mode. Two independent workflows Fig. 22 and Fig. 23 were used to deal with the missing data in Alteryx, and they were then saved separately to SQL Server. In Fig. 22 you can see the data being loaded in from the SQL server and run through lots of filters, one for each variable containing missing data, the filtered data is then written back to the SQL server. To impute the data Alteryx has a built in impute function as seen in Fig. 23. After deleting all outliers, there were 2102 observations with missing data, accounting for around 14.7% of the total data. To maintain the quality of the data, I opted to eliminate any data with missing values, justified by the fact that the data size would still be 11,762 and we would only need 300-500 according to guidelines, thus the data size requirement would still be met. Naturally, imputation would result in keeping more data which would also satisfy this requirement.

Fig. 23 Alteryx workflow to impute missing data

4.0 Analysis

4.1. SEM

As mentioned in the future work section of Student factors influencing STEM subject choice in Year 12 (Jeffries, Curtis and Conner, 2020) the school sector, parents' STEM status and greatest degree of education, and the child's self-efficacy have all been considered as factors that may impact a child's decision to pursue STEM in school. This information was not accessible at the time of this study, but I was able to find it in the most recent LSAY data collection. I felt that since a SEM model performed well for this study, that would be a good place to start with the new data.

Latent variable	Observed variables
MathScore	PV1MATH, PV2MATH, PV3MATH, PV4MATH, PV5MATH
SciScore	PV1SCIE, PV2SCIE, PV3SCIE, PV4SCIE, PV5SCIE
SelfEfficacy	ST67N01, ST67N02, ST67N03
MotivationSci	ST62N01, ST62N02, ST62N03, ST62N04

Table 5 Latent variable summary

After doing some investigation, I discovered that SPSS Amos was a structural equation modelling software and decided to utilize it to create the model. To begin, I opted to create four latent variables: MathScore, SciScore, SelfEfficacy, and MotivationSci, which I used to build a model along with the rest of the data in the data set to predict the outcome variable STEM. These latent variables are constructed using a combination of measured variables from the dataset, as shown in Table 5. In addition to these four latent factors, I included eight additional variables to predict the result variable; this hybrid SEM model combines observable and latent variables.

The model was created using Amos' built-in group or moderator function, to which gender was added so that both genders could be compared. Because independent variables are never perfect predictors of the dependent variable, every regression line has an error term. Instead, the line is a guess based on the data supplied. As a result, the error term indicates how confident you can be in the formula. Below you can see the initial model Fig. 24. The results of this model were saved before any changes were made.

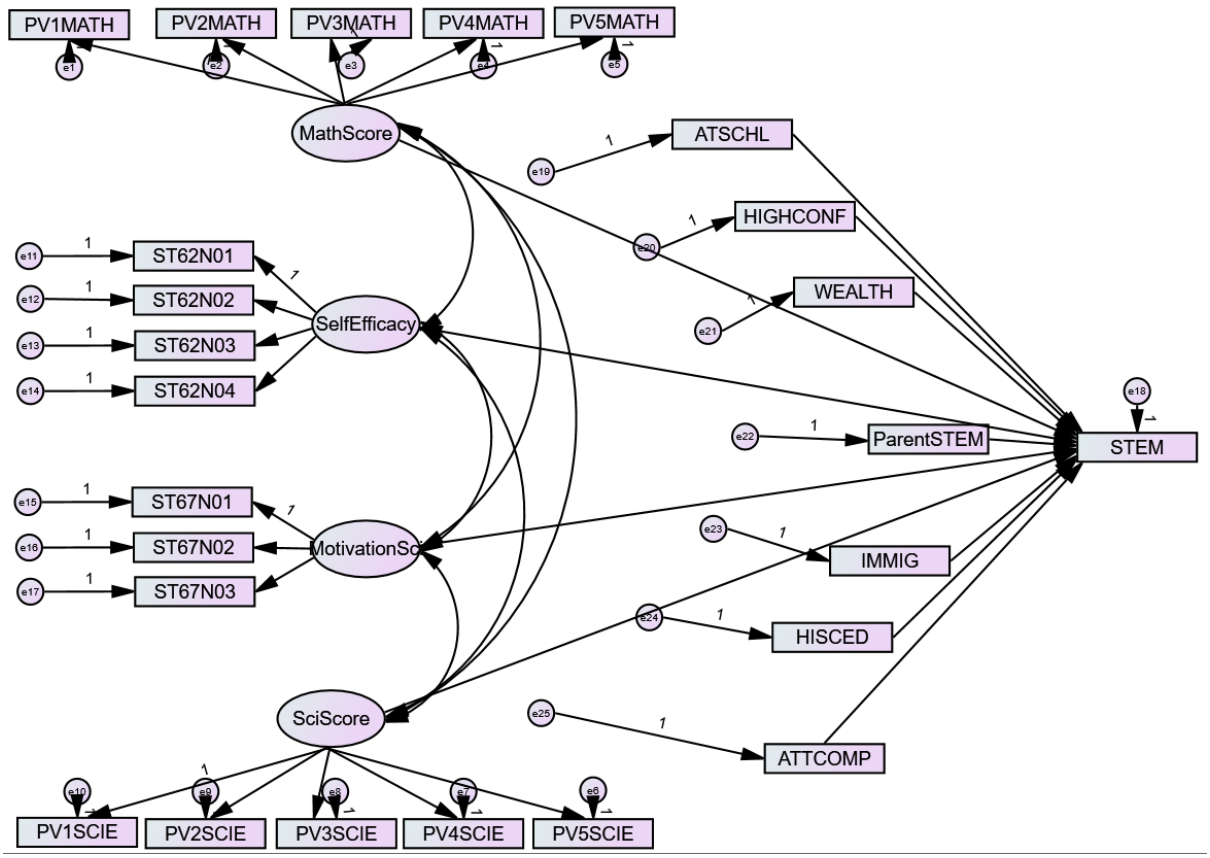


Fig. 24 Initial SEM model in SPSS Amos

I wanted to try optimizing the model after the first one was constructed, so I selected the modification indices in the analysis properties. This yielded a list of covariance and regressions that should be included in the model to improve fit. Because the modification indices suggested far too many modifications, the estimated parameter change threshold was increased from 4 to 50, and the model was run again. The list was now shorter, yet any modifications must be double-checked to ensure that they make logical sense before being applied to the model. There were still many adjustments to be made, making the model quite complex and difficult to interpret Fig. 25. The model was performed with both complete case data and imputed data, and the findings of each model were saved separately.

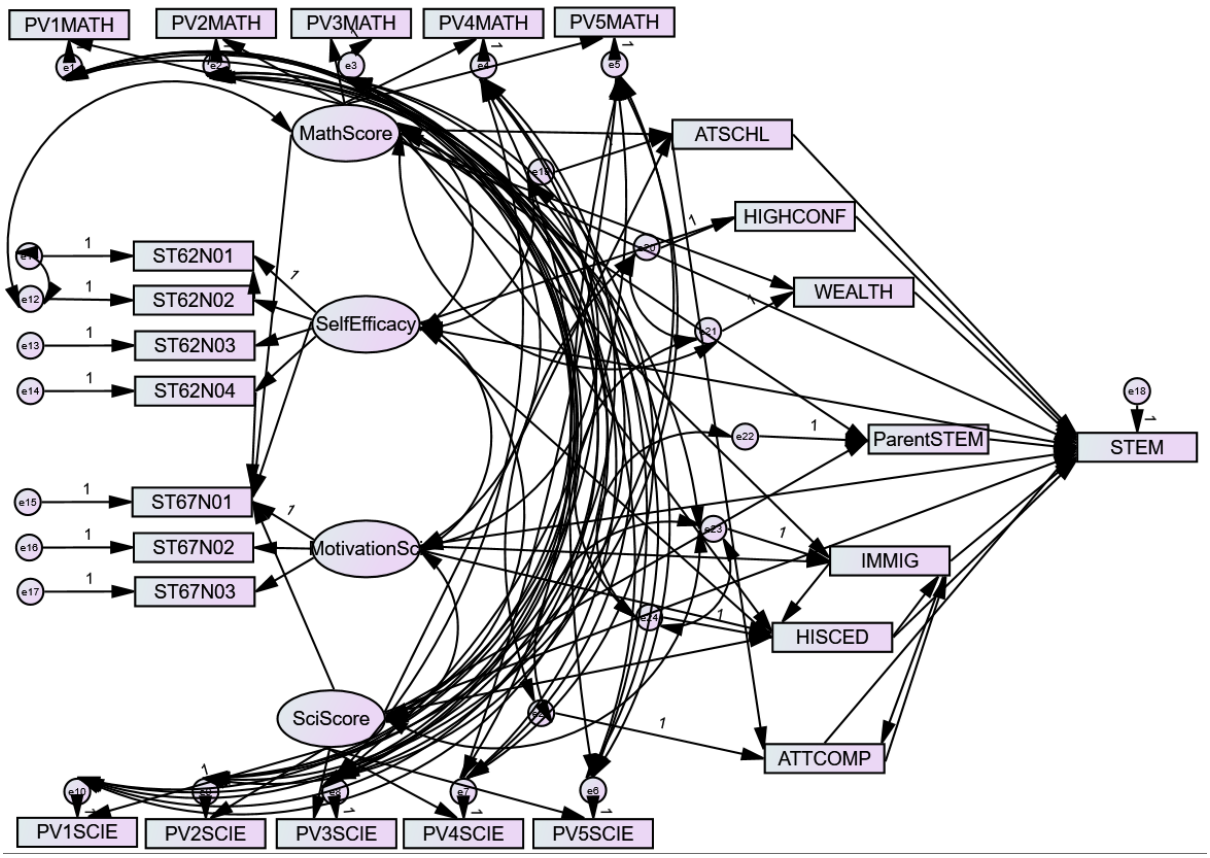


Fig. 25 Modified SEM model in SPSS Amos

Once I built my full SEM model in SPSS Amos, I decided to utilise the multiple group analysis function provided. This allowed me to perform a Chi-Square different test between the two groups. This will tell us if the two models are different based on the groups, although this is interesting, I can also create models with constrained paths to investigate isolated paths within the full model. In total I created 46 different models each one was constraining a different path within the model. SPSS Amos outputted a model comparison assuming the unconstrained model to be correct, it also gave us a P-value for each new model created, these values can be analysed to examine if any paths within the model are significantly different for the two groups.

R Packages used: lavaan, semPlot and caret.

Following the completion of the model in SPSS Amos, I decided to replicate it in R using the Lavaan package. Starting with the initial model created in Amos I created a model using the same covariance and regression in R as seen in Fig. 26. The model parameters are assigned to m1 and then the data is fitted using the lavaan Sem package with the gender variable specified as the group. I then obtained the model's fit measurements and saved them in an excel sheet.

```
m1 <- '
  # Latent vars
  MathScore =~ PV1MATH + PV2MATH + PV3MATH + PV4MATH + PV5MATH
  SciScore =~ PV1SCIE + PV2SCIE + PV3SCIE + PV4SCIE + PV5SCIE
  MotivationSci =~ ST67N01 + ST67N02 + ST67N03
  SelfEfficacy =~ ST62N01 + ST62N02 + ST62N03 + ST62N04

  STEM ~ ATSCHL + ATTCOMP + HISCED + HIGHCONF + IMMIG + MathScore +
MotivationSci + ParentSTEM + SciScore + SelfEfficacy

 # covariance
    MathScore ~~ SelfEfficacy
    MathScore ~~ MotivationSci
    MathScore ~~ SciScore
    SelfEfficacy ~~ MotivationSci
    SelfEfficacy ~~ SciScore
    MotivationSci ~~ SciScore
'

fit1 <- lavaan::sem(m1, data=LSAYScale, group = "ST04Q01")
```

Fig. 26 R code for SEM model

The more complex updated model produced in Amos was also replicated in R and the findings were recorded. To determine if the model could be simplified by removing some of the variables Fig. 27, I took a top-down approach, removing each variable one at a time and saving the outputs. I then experimented with various combinations of the improved models. When the best models were found, they were put to the test with both sets of data and the results were kept.

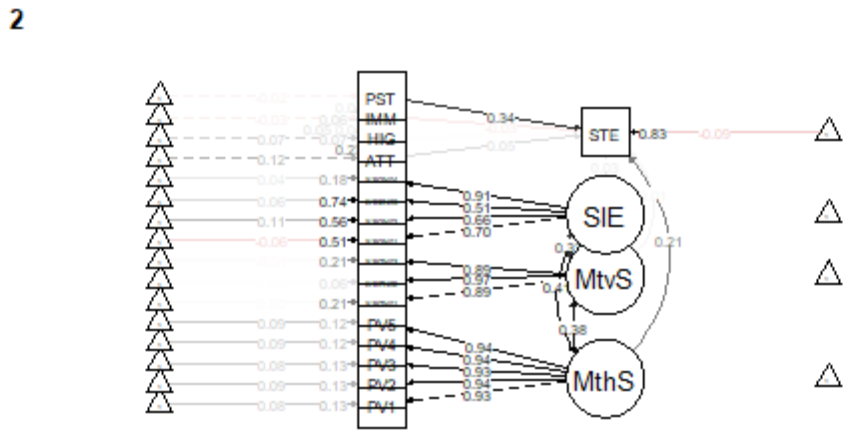


Fig. 27 Simplified SEM model in R

4.2. Logistic Regression

R Packages used: tidyverse, caret and dlookr.

Multinomial Logistic Regression is a classification approach used in statistics to estimate the probabilities of multiple possible outcomes of a categorically distributed dependent variable given a set of independent factors. I wanted to see if I could make predictions after finishing the SEM model, this is the first of four different supervised machine learning algorithms used to build models in an attempt to make predictions. I'm interested if we can successfully predict whether a student would choose a STEM subject based on the variables SECTOR, ST04Q01, IMMIG, HISCED, ATTCOMP, ATSCHL, HIGHCONF, WEALTH, parentSTEM, MathScore, SciScore, SelfEfficacy, and MotivationSci. Since the data has a binary outcome, Multiple Logistic Regression was chosen over Multiple Linear Regression. I did consider using Jamovi for the Logistic Regression however it is limited to 10,000 rows of data, and I have over 14,000, therefor I completed this in R.

```
# Create new variables ----------------------------------------------------
LSAY$MathScore <- (LSAY$PV1MATH + LSAY$PV2MATH + LSAY$PV3MATH + LSAY$PV4MATH +
LSAY$PV5MATH)/5
LSAY$SciScore <- (LSAY$PV1SCIE + LSAY$PV2SCIE + LSAY$PV3SCIE + LSAY$PV4SCIE +
LSAY$PV5SCIE)/5
LSAY$SelfEfficacy <- (LSAY$ST62N01 + LSAY$ST62N02 + LSAY$ST62N03 +LSAY$ST62N04
)/4
LSAY$MotivationSci <- (LSAY$ST67N01 + LSAY$ST67N02 + LSAY$ST67N03)/3
```

Fig. 28 Creating new variables in R

In this case, I just find the mean of the predictor variables to produce the latent variables, as shown in Fig. 28. To evaluate the model, I separated the data into testing and training sets. The data was then randomly split into an 80/20 proportion for the train and test datasets Fig. 29, yielding 8,337 observations for training and 2,084 for evaluating the model for the complete case data and 11,092 observations for training and 2,772 for evaluating the model for the imputed data, in both cases, using a seed of 123.

```
# Split data into training and test set ----------------------------------
set.seed(123)
training.samples <- LSAY$STEM %>%
  createDataPartition(p = 0.8, list = FALSE)
train1 <- LSAY[training.samples, ]
test1 <- LSAY[-training.samples, ]

training.samples <- LSAYImpute$STEM %>%
  createDataPartition(p = 0.8, list = FALSE)
train2 <- LSAYImpute[training.samples, ]
test2 <- LSAYImpute[-training.samples, ]
```

Fig. 29 Creating test and train split in R

First, I used the glm function to fit all variables into a generalized linear model, and so the family was set to binomial to describe the binary result Fig. 30.

```
# Build model ----------------------------------------------------------------
model1 <- glm( STEM ~ SECTOR + ST04Q01 + IMMIG + HISCED + ATTCOMP + ATSCHL +
HIGHCONF + WEALTH + ParentSTEM + MathScore + SciScore + SelfEfficacy +
MotivationSci, data = train1, family = binomial(link='logit'))
summary(model1)
```

Fig. 30 Logistic Regression model in R

The beta coefficient estimations and their significance levels may then be seen by looking at the output coefficients table Fig. 31. HISCED, HIGHCONF, and SCISCORE were removed to form a second model because they were not significant in this case. Leaving these statistically insignificant variables in the model may result in overfitting.

```
Call:
glm(formula = STEM ~ SECTOR + ST04Q01 + IMMIG + HISCED + ATTCOMP +
    ATSCHL + HIGHCONF + WEALTH + ParentSTEM + MathScore + SciScore +
    SelfEfficacy + MotivationSci, family = binomial(link = "logit"),
    data = train1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8580  -1.0430   0.3616   0.9830   2.0064

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.9041172  0.2098233  -9.075  < 2e-16 ***
SECTOR        -0.1980987  0.0404044  -4.903 9.44e-07 ***
ST04Q01       -0.4490401  0.0518389  -8.662  < 2e-16 ***
IMMIG         -0.1155158  0.0417942  -2.764  0.00571 **
HISCED        -0.0599022  0.0230021  -2.604  0.00921 **
ATTCOMP        0.0747439  0.0294751   2.536  0.01122 *
ATSCHL         0.1433546  0.0259344   5.528 3.25e-08 ***
HIGHCONF       0.0047742  0.0286646   0.167  0.86772
WEALTH         0.0981492  0.0325971   3.011  0.00260 **
ParentSTEM     2.3548186  0.0925581  25.442  < 2e-16 ***
MathScore      0.0051439  0.0008634   5.958 2.56e-09 ***
SciScore       0.0006909  0.0007833   0.882  0.37774
SelfEfficacy   0.0782730  0.0369977   2.116  0.03438 *
MotivationSci  0.0472273  0.0291619   1.619  0.10534
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11103.4  on 8336  degrees of freedom
Residual deviance:  9261.1  on 8323  degrees of freedom
AIC: 9289.1

Number of Fisher Scoring iterations: 5
```

Fig. 31 Summary of full Logistic Regression model in R

The test data was used to evaluate all models by predicting the outcome variable with R's predict function Fig. 32. This predicts the probability of a student doing STEM; to evaluate, it needed to be converted to the same structure as the actual variable. To begin, the probability was transformed to a 0 or a 1 using an if-else statement that specifies that if the likelihood is more than 50%, it should be assigned 1 and if not, it should be assigned 0. These were then converted to factor types and compared to the test dataset's real data. The expected and actual outcome variables were entered into a confusion matrix, which was then saved.

```
# Predict ------------------------------------------------------------------
prob1 <- model1 %>% predict(test1, type = "response")
predicted1 <- ifelse(prob1 > 0.5, "1", "0")
predicted1 <- as.factor(predicted1)
table(predicted1, test1$STEM)
actual1 <- as.factor(test1$STEM)
confusionMatrix(predicted1, actual1)
mean(predicted1 == test1$STEM)
```

Fig. 32 Predicting using Logistic Regression in R

4.3. Naïve Bayes

R Packages used: rsample, dplyr, ggplot2, caret and e1071.

Naïve Bayes is a supervised machine learning technique that uses the Bayes Theorem to tackle a variety of classification problems. To use this strategy, I began by constructing new variables Fig. 28 and a test and train dataset Fig. 29 in the same way that I did with the Logistic Regression approach. The test and train sets, however, had to be set up slightly differently for the Naïve Bayes algorithm. This split similarly employed a seed of 123 and a split of 80/20, but the test dataset needed to contain all of the data needed to predict the outcome variable, omitting the actual outcome variable, as denoted by the -9. (STEM being the 9th column). The 9th column then is saved in a separate list of factors labelled truth1 and truth2 for each of the two data sets, which will be used to evaluate the model Fig. 33.

```
# Train and test model -----------------------------------------------------
set.seed(123)
# Create testing and training data sets
sample_size1 <- floor(0.8*nrow(LSAY))
train_indices1 <- sample(seq_len(nrow(LSAY)),size = sample_size1)
train1 <- LSAY[train_indices1,]
test1 <- LSAY[-train_indices1,-9]

# Save for evaluation
truth1 <- LSAY[-train_indices1,9]
truth2 <- LSAYImpute[-train_indices2,9]
```

Fig. 33 Naïve Bayes test and train data

The model is then saved once it has been fitted to the train data with the naiveBayes function Fig. 34. The fits are then used to produce predictions on the test data, which are contained in two lists titled fit1 and fit2. The projected values are then compared to the actual values in a confusion matrix.

```
# Fit model
fit1 <- naiveBayes(STEM ~ ., data=train1)
fit2 <- naiveBayes(STEM ~ ., data=train2)

predict1 <- predict(fit1,test1)
predict2 <- predict(fit1,test2)

confusionMatrix(predict1,truth1)
confusionMatrix(predict2,truth2)
```

Fig. 34 Naïve Bayes predictions

I decided to tune the model to see if it could be improved upon. This entailed creating a tuning grid, resampling, and determining the optimal model. The nb_grid tuning grid has predefined parameters to test the model with, in order to discover the best settings to apply to it. This includes determining whether the model should employ a kernel density estimate, adjusting the kernel density value's bandwidth, and Laplace smoothing parameters Fig. 35. After this the models were fitted using this stated tuning grid, which took a considerable time to complete. The final tuning parameters for the models could then be displayed, and the outcomes of these freshly tuned models could be saved as well.

```
# Tuning grid  ------------------------------------------------------------
# Define
nb_grid <-   expand.grid(usekernel = c(TRUE, FALSE),
                         laplace = c(0, 0.5, 1),
                         adjust = c(0.75, 1, 1.25, 1.5))

# Fit model
fitTuned1 <- train(STEM ~ ., data = train1,
                                method = "naive_bayes",
                                usepoisson = TRUE,
                                tuneGrid = nb_grid)

fitTuned2 <- train(STEM ~ ., data = train2,
                   method = "naive_bayes",
                   usepoisson = TRUE,
                   tuneGrid = nb_grid)

# Selected tuning parameters
fitTuned1$finalModel$tuneValue
# 0     FALSE   0.75
fitTuned2$finalModel$tuneValue
# 0     FALSE   0.75
```

Fig. 35 Naïve Bayes tuning

4.4. Decision Tree

R Packages used: C50, rpart, rpart.plot, caret, ROCR, adabag, rattle and gbm.

By learning simple decision rules derived from data attributes, decision trees generate a model that predicts the value of a target variable. Like the two approaches above, this began with loading the data in from the SQL server creating the new variables, casting them into factors and partitioning the data into a training and testing dataset. The data was split using the same approach as seen in Fig. 29. I wanted to try two different methods so completed the model using both the C50 and Rpart algorithms.

```
# Create a decision tree model using C5.0 --------------------------------
C5.0Control <- C5.0Control(
  subset = FALSE,
  bands = 0,
  winnow = FALSE,
  noGlobalPruning = FALSE,
  CF = 0.25,
  minCases = 2,
  fuzzyThreshold = FALSE,
  earlyStopping = TRUE
)
```

Fig. 36 Decision Tree model using C5.0

Once the data was ready, I create a Decision Tree model using C5.0 Fig. 36. The C5.0 control method defines the control aspects of the C5.0 fit, the first parameter we provide to the method is a boolean subset which indicates whether the model must examine groups of discrete predictors for splits. The next is an int value called bands ordering the groups into specified bands, I have set this value to 0 as it is not relevant to the model. The winnow variable again is another boolean which denotes the use of feature selection, the descriptive noGlobalPruning attribute toggles the use of a final pruning step to simplify the tree. The next aspect we specify is the confidence factor value, which is a number between 0 and 1, the minCases variable decides the minimum amount of samples that must be allocated to at least two splits. FuzzyThreshold and earlystopping are the final toggles we specify, FuzzyThreshold controls whether to utilise advance splits of data, earlyStopping allows us to elect to use the internal method for stopping boosting or not. The model was then applied to the test data to make predictions and was evaluated using a confusion matrix. Multiple different combinations of these settings were tried but the selected parameters seemed to work the best.

```
# Create model using rpart ------------------------------------------------
rpart.Control <- rpart.control(
  minsplit = 20,
  xval = 10
)
```

Fig. 37 Decision Tree model using Rpart

I then moved on to create the same model using the Rpart function as seen in Fig. 37. The rpart.control method defines the controlling aspects of the 'rpart' fit. Two parameters I have utilised are minsplit and xval, the minsplit variable is the minimum number of observations that must exist in a particular node for a split to be attempted. The xval parameter selects the number of cross-validations used in the fit. This model was tested in the same way. When looking at the results I noticed the two algorithms gave identical results for the imputed data, so I tried to get the model to improve by adding a bagging model too. Bagging, short for Bootstrap aggregating, is a machine learning ensemble meta-algorithm that aims to increase the accuracy and robustness of machine learning algorithms used in statistical classification. The bagging model was the simplest to implement of the three methods Fig. 38, once this was completed, I made a prediction from which I evaluated and saved the results.

```
# Bagging model ----------------------------------------------------------
fit5 <- bagging(
  STEM ~ .,
  data = train1
)
```

Fig. 38 Decision Tree model using bagging

4.5. Random Forest

R Packages used: MASS, rpart, rpart.plot, adabag, randomForest, ggplot2 and caret.

A Random Forest is a machine learning technique for solving classification and regression problems. It makes use of ensemble learning, which is a technique for solving complicated problems by combining several classifiers. Some of the data for the Random Forest method had an excessive number of factors and needed to be binned Fig. 39. To accomplish so, I used the cut function to specify the number of breaks by taking the minimum and maximum values of the relevant columns and rounding up the number of steps in the data.

```
# Bin vars with large amount of factors
LSAY$ATSCHL <- cut(LSAY$ATSCHL, breaks = 5, labels = c("1", "2", "3", "4", "5"))
LSAY$HIGHCONF <- cut(LSAY$HIGHCONF, breaks = 6, labels = c("1", "2", "3", "4",
"5", "6"))
LSAY$WEALTH <- cut(LSAY$WEALTH, breaks = 6, labels = c("1", "2", "3", "4", "5",
"6"))

LSAYImpute$ATSCHL <- cut(LSAYImpute$ATSCHL, breaks = 5, labels = c("1", "2", "3"
, "4", "5"))
LSAYImpute$HIGHCONF <- cut(LSAYImpute$HIGHCONF, breaks = 6, labels = c("1", "2",
"3", "4", "5", "6"))
LSAYImpute$WEALTH <- cut(LSAYImpute$WEALTH, breaks = 6, labels = c("1", "2", "3"
, "4", "5", "6"))
```

Fig. 39 Binning variables in R

The randomForest function in R was then used to fit the model, this was then used to create predictions, which were then evaluated. I used parameter tuning to try to improve the model. One search approach you might do is to test random values within a range. This is useful since we are unsure of the value and wish to eliminate any biases we might have when setting the parameter. The search process is slowed by both 10-fold cross-validation and 3 repeats Fig. 40, however, this is done to limit and reduce overfitting on the training set. This took a long time to finish running.

```
# Random Search
control <- trainControl(method="repeatedcv", number=10, repeats=3, search
="random")

mtry1 <- sqrt(ncol(train1))
fit5 <- train(STEM ~., data=train1, method="rf", tuneLength=15, trControl
=control)
print(fit5)
plot(fit5)
```

Fig. 40 Random search tuning

The complete cases were optimized at 10 mtry Fig. 40, and the imputed data at 20 mtry Fig. 41, as a result of the tuning. At each split, mtry is the number of variables randomly picked as possibilities. The outcomes of these optimized models were saved, and then all models were compared and contrasted.

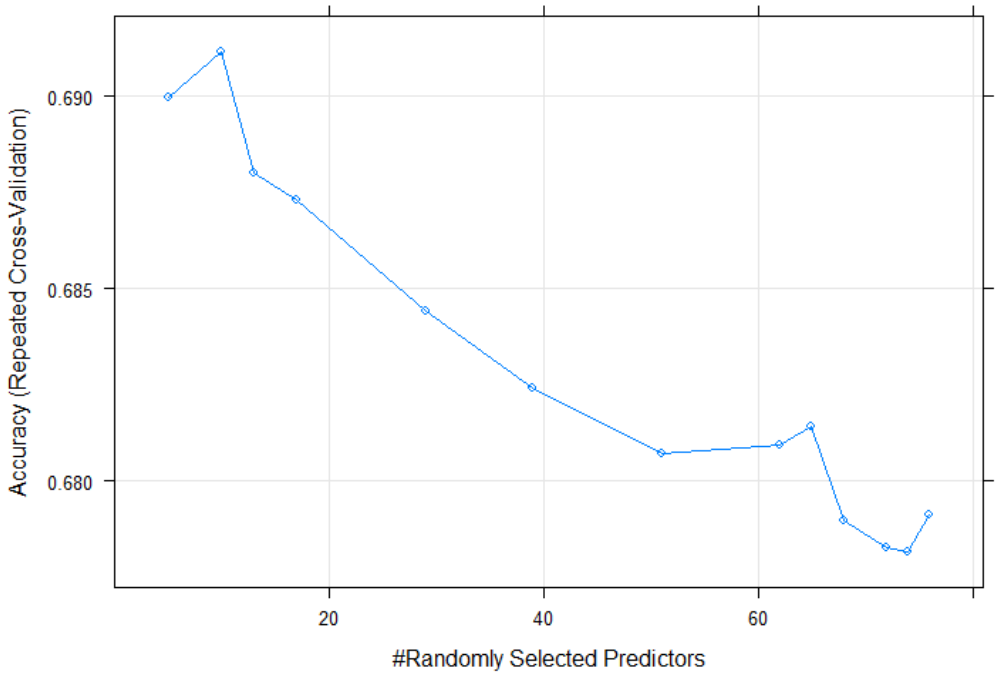


Fig. 41 Complete cases randomly selected predictors

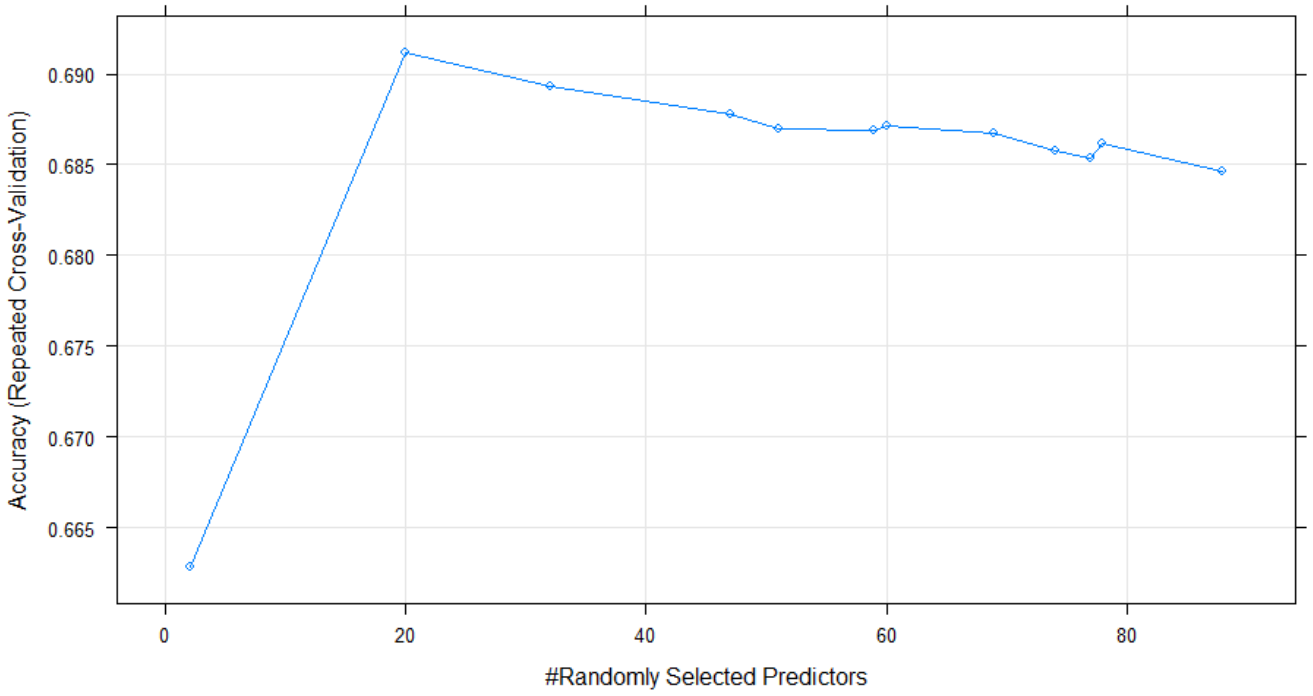


Fig. 42 Imputed randomly selected predictors

5.0 Results

5.1. Analysis

During the preliminary analysis, it was found that although the uptake of STEM degrees was consistently rising, the disparity between men and women still exists Fig. 10. Using the forecast tool built into Tableau, with exponential smoothing we were able to gain insights into the projections for the number of degrees that will be conferred in 2019 Fig. 43. This data is available from the same source therefore I could compare my prediction against the existing data for the same year. We found that the total number of degrees was forecasted to be 414,988, an over-estimation by 5.07% compared to the actual figures 412,894. We had speculated an increase of 5% from the previous year when we actually found there to be a 4.47% increase, the forecast had overestimated the male cohort by 1.07% and it had underestimated the females by 4.38%.

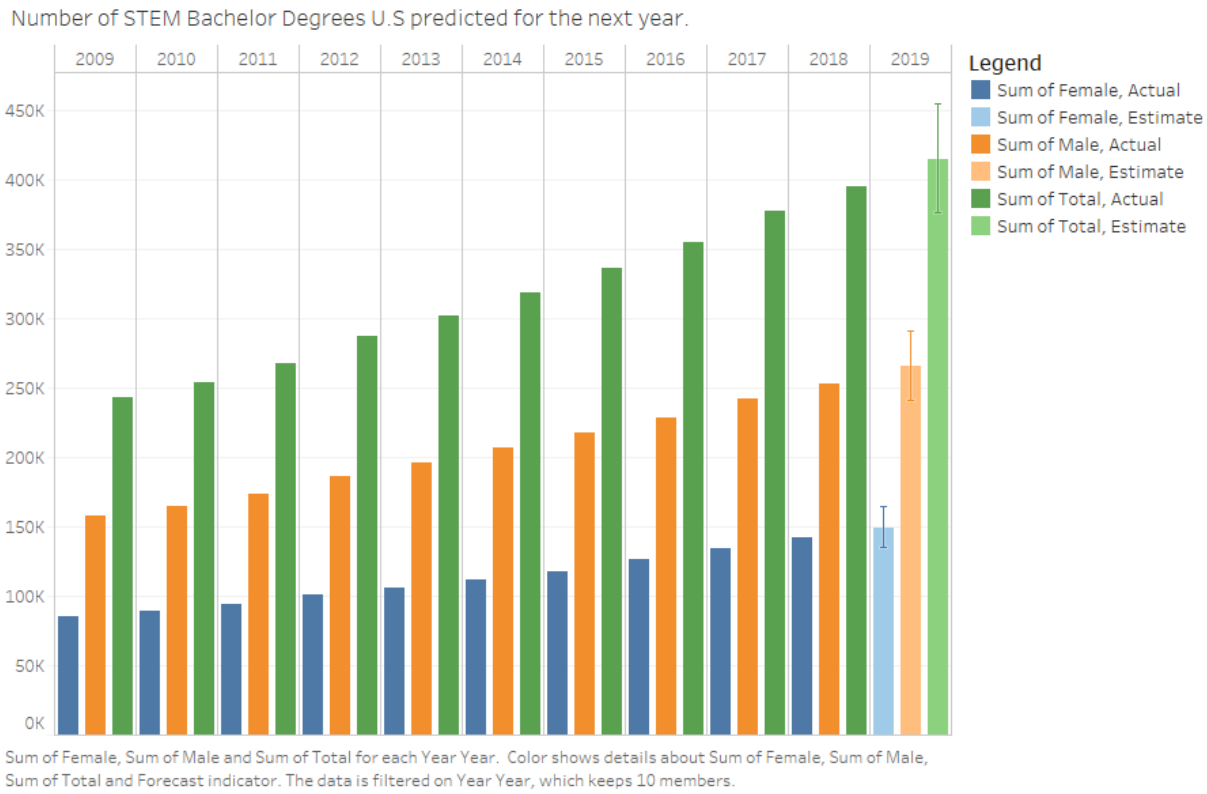


Fig. 43 Prediction of bachelor's degrees

The pew dataset provided some interesting insights mostly supplied by those who work in a STEM-related field. These are best represented by graphical visualizations in order to examine the data clearly and coherently, the results are discussed below.

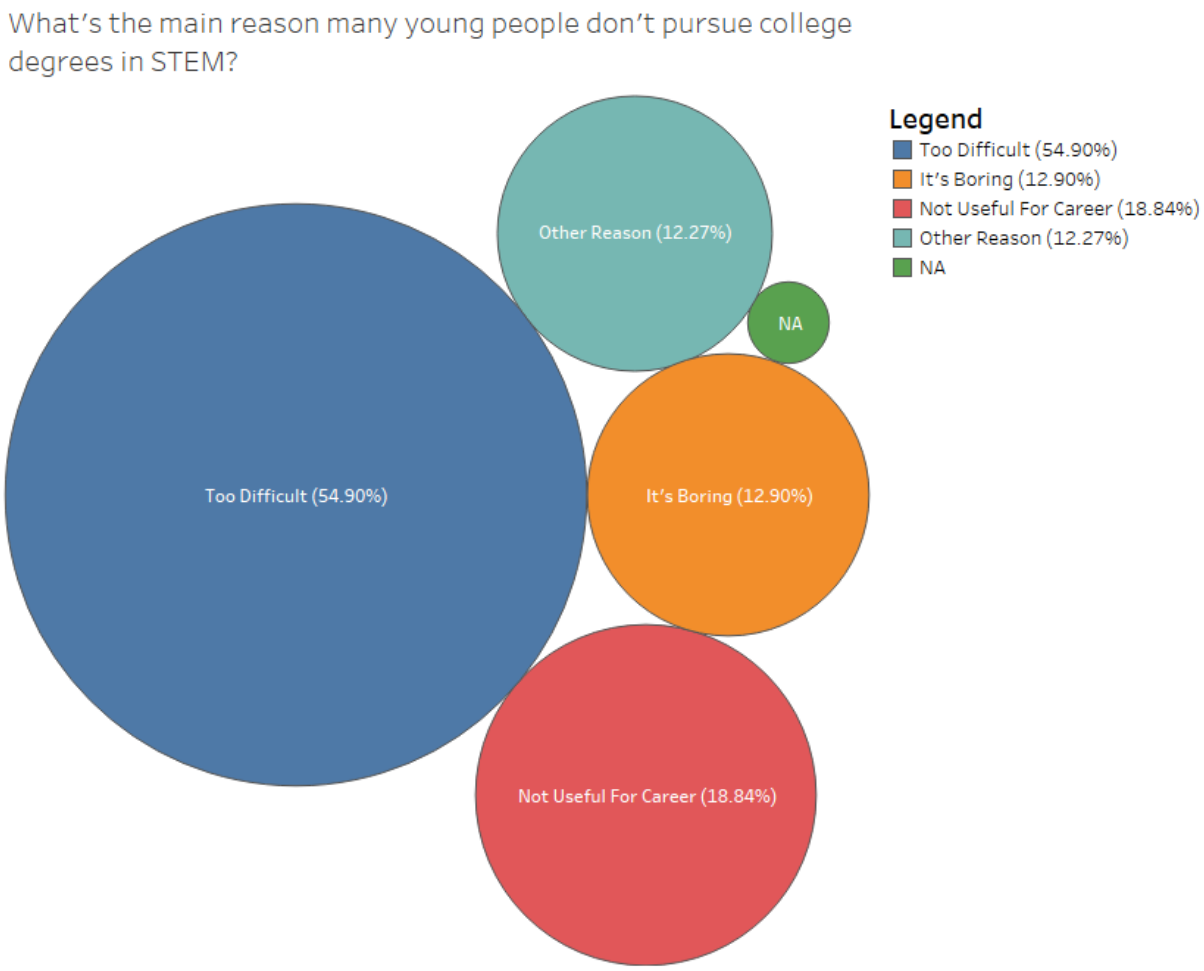


Fig. 44 Reasons for not pursuing STEM

As seen in Fig. 44, over half of the participants surveyed believed that the main reason that many young people do not pursue STEM degrees is that they find the subject matter too difficult. Nearly a fifth of all those surveyed claimed that they found STEM was not relevant to their career prospects and is deemed not useful. Just under 13% of respondents found STEM and STEM-related fields to be boring and did not catch the interest of the participant. Around 12% of those surveyed cited various reasons of little significance as to why they did not pursue a career in STEM. A mere 1% of respondents elected to not answer the question. Difficulty appears to be the prevailing deterrent for pursuing STEM, various potential factors can influence this such as outdated teaching methodologies, low self-confidence and preconceived notion that STEM is reserved for those of high academic competency. The net prevailing issue is that some respondents felt that the STEM degree was not useful for their choice of career, this suggests that those who answered possibly aspire to achieve a non-STEM(Arts) degree and therefore do not feel STEM is relevant to them. The last primary reason that young people elected not to choose is that they found it boring, this suggests a fundamental lack of interest which can stem from a lack of role models in the field, lack of exposure or it also could be that the respondent has a higher interest in another topic.

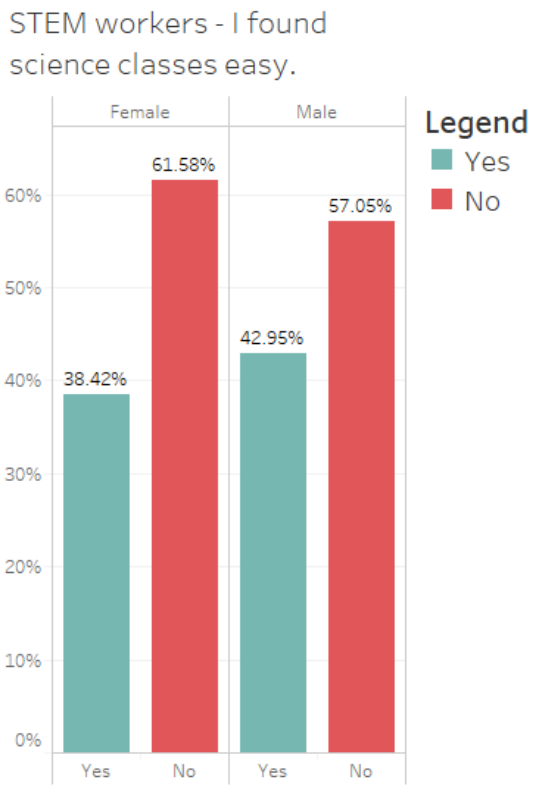


Fig. 45 Sch10A 1 summary

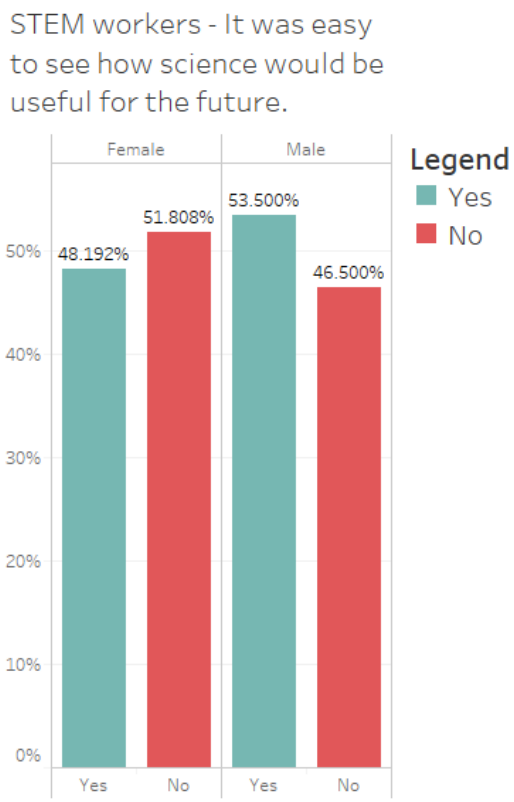
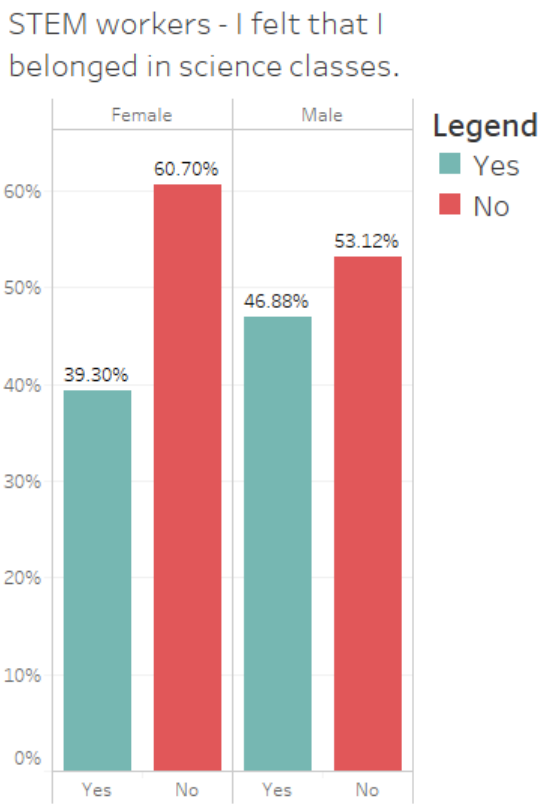


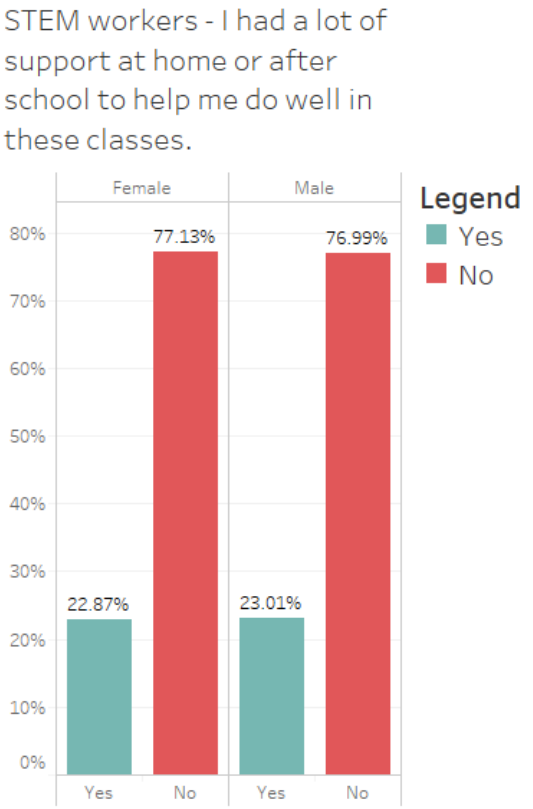
Fig. 46 Sch10A 2 summary

I investigated the consensus amongst those who work in STEM and how they felt about their experience learning science topics during early education. The majority of workers of both genders found Science in school not to be easy Fig. 45, males accounted for 4% more of the yes cohort than females. These results counteract the notion that students cannot do STEM if they find it difficult, a large proportion of people still pursue these careers despite not finding them easy. Males and females are nearly evenly divided on the issue of whether they can see how science can be useful for the future Fig. 46, this suggests that students are not being taught the practical application of the use of science in everyday life and its potential as a career, another potential cause is a lack of exposure to science from an academic perspective and a lack of understanding can create a disparity on the value of science as a whole.



% of Total Count of Case ID for each Sch10A 3 broken down by Ppgender. Color shows details about Sch10A 3. The data is filtered on Worktype Final, which keeps 1. The view is filtered on Sch10A 3, which keeps No and Yes. Percents are based on each pane of the table.

Fig. 47 Sch10A 3 summary



% of Total Count of Case ID for each Sch10A 5 broken down by Ppgender. Color shows details about Sch10A 5. The data is filtered on Worktype Final, which keeps 1. The view is filtered on Sch10A 5, which keeps No and Yes. Percents are based on each pane of the table.

Fig. 48 Sch10A 5 summary

When asked whether they felt that they belonged in a science class Fig. 47, female respondents felt significantly less so than their male counterparts. Male students were nearly evenly split on the issue with 46.88% claiming they did not belong although females were noticeably lower at 39.3%, this is a very significant deterrent as those who feel they do not belong in the class might also feel they do not belong in the industry, it difficult to determine what the exact some reason student might feel they don't belong, it is possibly a result of low self-efficacy or even a skewed distribution of males to females in a classroom. On the issue of whether respondents felt they had sufficient support at home Fig. 48, both males and females felt they were not supported in equal amounts with both genders scoring in or around 23%. The lack of support from home I believe is a result of student parents falling into the same trap where they were not supported during their time at school, this lack of support can result from parents not being able to aid their children with technical homework or assignments which contain advanced mathematic or scientific concepts. Many people who do not use or practise these concepts in their day-to-day careers will not be as competent as they have not studied the topics since their time in school.

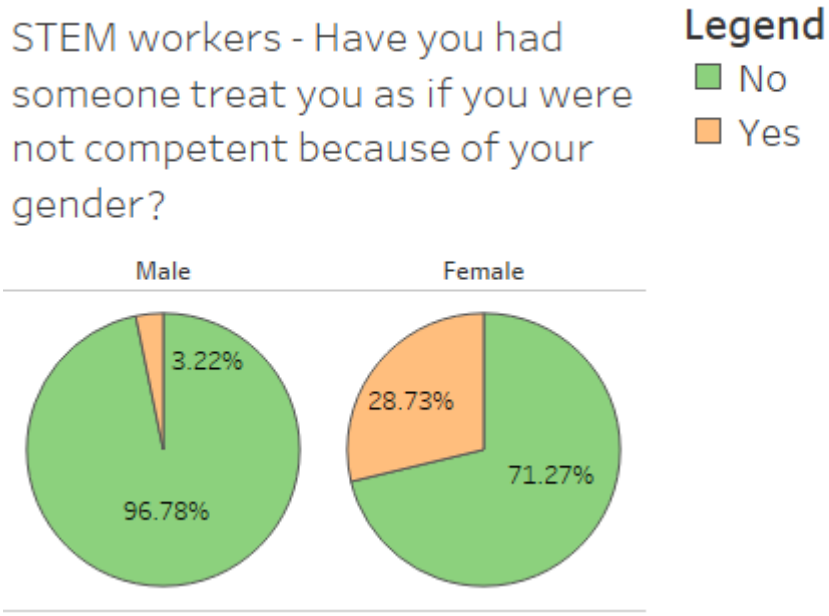


Fig. 49 Not competent due to gender summary

Perhaps the most significant insight is seen in Fig. 49, when asked whether the participant had ever been treated as incompetent as a result of their gender, female STEM workers answered yes nearly 10 times the amount that of male ones, a mere 3.22% of men had answered yes compared to 28.73% of the women asked. This reveals a striking difference between the cohorts and could be a major deterrent for women who are considering working in STEM.

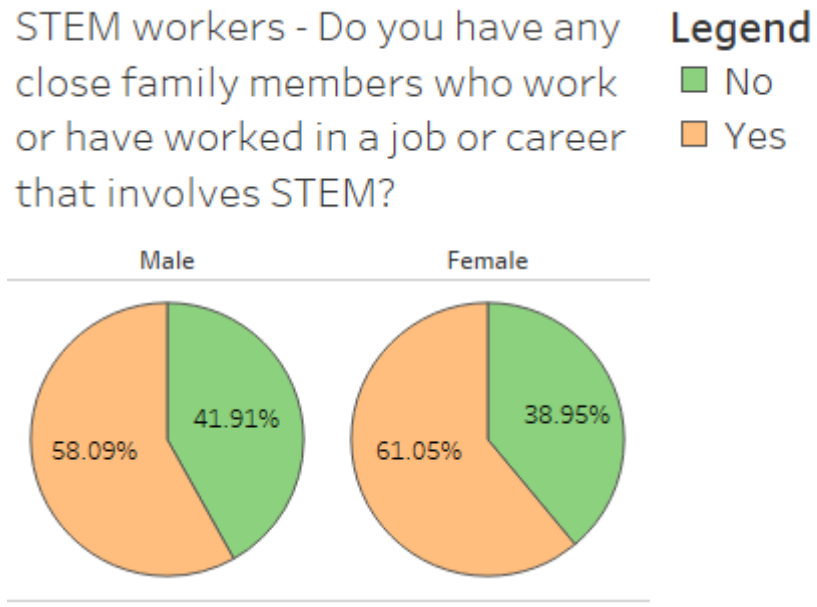


Fig. 50 Family in STEM summary

The number of people with families who work in STEM is relatively the same amongst men and women Fig. 50, with females edging at 61.05% compared to males with 58.09%. In both cases, over half the respondents asked claimed they had family members who worked in STEM fields. This could possibly influence them to take up STEM by being exposed to the industry as well as having a potential role model in their family.

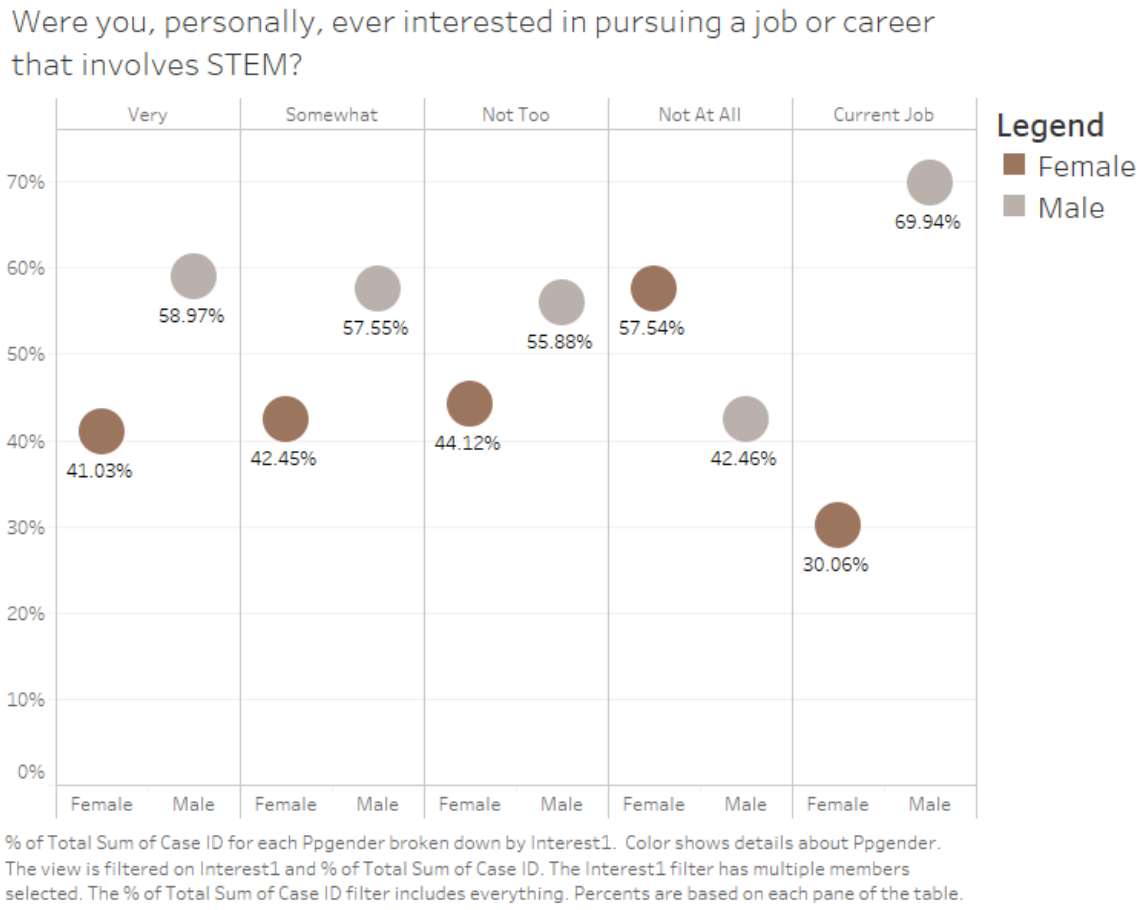


Fig. 51 Interest in pursuing STEM summary

When the full population sample was asked if they were ever interested in pursuing a career in STEM, 69.94%of those who stated it was their current occupation were men, highlighting an imbalance when compared to only 30.06% of women Fig. 51. When claiming to have no interest at all in a STEM career, women occupied most of the answers at 57.54%. This is the only female-dominated preference of the four levels of interest specified this answer also has the lowest male representation of all the answers.

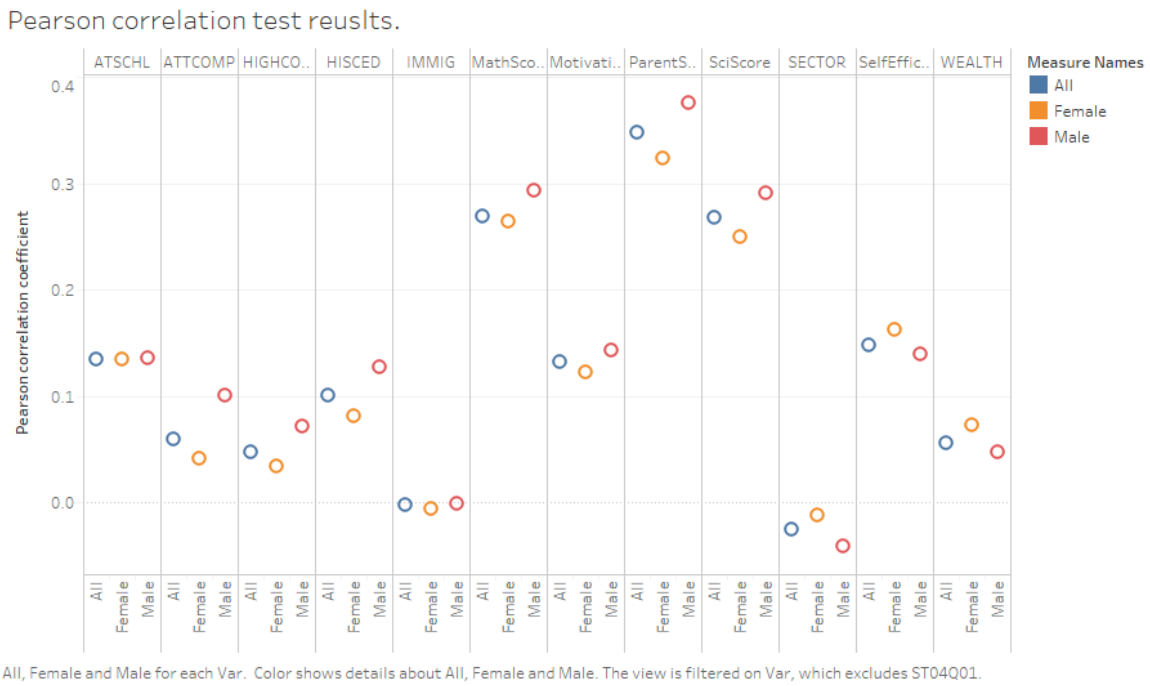


Fig. 52 Pearson correlation test summary

Lastly, I completed a Pearson correlation test in SPSS using their built-in function. The hypothesis for this test is:

H0: r = 0

H1: r ≠ 0

The results were saved into an excel sheet which was used to make a graph in Tableau. All variables in the data that were going to be used in the modelling stage were tested to see if they were correlated with the outcome variable STEM. As seen in Fig. 52, ParentSTEM is positively correlated with the outcome variable, this is the highest correlated variable for both males and females. It is the only correlation that's moderately correlated to the STEM variable. MathScore and SciScore are also positively correlated although not to the same extent as ParentSTEM. Although small, ATSCHL, HISCED, MotivationSCi and SelfEfficacy are minor positive correlations. ATTCOMP, HIGHCONF and WEALTH have very little correlations with IMMIG having nearly none at all. The only negatively correlated variable is SECTOR although this correlation is very weak. Overall, we can reject H0 in favour of H1 although IMMIG only has a correlation coefficient of -0.002 it is still not equal to 0.

5.2. SEM Model

As mentioned in the section above, two SEM models were built in SPSS Amos Fig. 24 and Fig. 25. The results of these models are shown in Table 6. The evaluation metrics used to validate SEM models are Chi-Squared statistics, GFI, AGFI, NFI, TLI, CFI and RMSEA. Although chi-square statistics are presented, they are not utilized to evaluate model fit because

models may fail a chi-square test simply because of the huge sample size (Schermelleh-Engel, Moosbrugger and Müller, 2003). The goodness of fit index (GFI) is a metric for how well the hypothesized model and the observed covariance matrix fit together. The adjusted goodness of fit index (AGFI) corrects the GFI, which is influenced by the number of latent variable indicators. For GFI and AGI the closer the value is to one the better with values over 0.9 suggesting a good fitting model (Hooper, Coughlan and Mullen, 2007). Normed Fit Index (NFI) also specifies a cut-off point for an optimal fitting model of 0.9 with more recent studies suggesting this should be raised to 0.95. This statistic evaluates the model by comparing the model's chi-squared value to the null model's chi-squared statistic. CFI is a revised form of the NFI which takes into account sample size, this should also have a cut-off point of 0.95. With continuous data, the Tucker–Lewis index (TLI) heavily relies on the usual cut-off values defined under normal-theory maximum likelihood (ML), which should be above 0.90. The RMSEA is the last evaluation metric, which shows us how well the model fits the population covariance matrix with unknown but ideally chosen parameter estimates. For the RMSEA in a well-fitting model the lower limit is close to 0 while the upper limit should be less than 0.08.

	M1 Complete	M1 Imputed	M2 Complete	M2 Imputed
Chi-Square	22491.149	28578.4	1032.165	1413.6
P-value	0	0	0	0
GFI	0.863	0.872	0.992	0.992
AGFI	0.833	0.843	0.988	0.987
NFI	0.905	0.913	0.996	0.996
TLI	0.895	0.904	0.996	0.996
CFI	0.907	0.915	0.997	0.996
RMSEA	0.063	0.062	0.012	0.013

Table 6 SEM evaluation metrics SPSS Amos

As seen in the above table the first model Fig. 24 Initial SEM model in SPSS Amos, does not meet the specified criteria for an acceptable model, the GFI, AGFI, NFI and TLI were below the accepted values, therefore, this model is evaluated as a poor fitting model. However, the second modified model Fig. 25 Modified SEM model in SPSS Amos met all our evaluation criteria and is accepted as a good fitting model. For the first model, the imputed data provided superior results, however in the second model, the complete case data gave us better results, there is a marginal difference overall between the two data sets. This does not indicate that the complete cases or imputed data work any better with the models. The modified model with complete case data is the best fitting model, due to all of the modifications the results can be difficult to read from the graph and are therefore presented in Table 7 for each group (male and female).

Standardized regression weights	Male	Female
ATSCHL ← MathScore	0.135	0.152
ATSCHL ← SciScore	0.154	0.152
HISCED ← SciScore	0.226	0.261
HISCED ← MathScore	0.225	0.209
HISCED ← MotivationSci	-0.018	-0.079
HISCED ← SelfEfficacy	-0.013	-0.071
IMMIG ← MotivationSci	0.077	0.07
ATTCOMP ← ATSCHL	0.087	0.062
IMMIG ← MathScore	0.063	0.127
ST62N02 ← SelfEfficacy	0.688	0.67
ST62N04 ← SelfEfficacy	0.874	0.836
HIGHCONF ← SelfEfficacy	0.198	0.115
ParentSTEM ← SciScore	0.137	0.126
ParentSTEM ← MathScore	0.135	0.132
WEALTH ← MathScore	0.074	0.084
ST62N01 ← SelfEfficacy	0.91	1.301
ST62N03 ← SelfEfficacy	0.517	0.481
ST67N01 ← MotivationSci	0.823	0.84
ST67N02 ← MotivationSci	0.976	0.984
ST67N03 ← MotivationSci	0.887	0.907
STEM ← ParentSTEM	0.31	0.278
STEM ← IMMIG	-0.028	-0.025
STEM ← MathScore	0.182	0.175
STEM ← SelfEfficacy	-0.022	-0.018
STEM ← MotivationSci	-0.021	0.002
STEM ← HISCED	-0.038	-0.081
STEM ← WEALTH	0.022	0.036
STEM ← HIGHCONF	-0.008	0.001
STEM ← ATSCHL	0.048	0.05
PV3MATH ← MathScore	0.934	0.928
PV4MATH ← MathScore	0.937	0.93
PV5MATH ← MathScore	0.936	0.926
PV2SCIE ← SciScore	0.948	0.941
PV1SCIE ← SciScore	0.949	0.939
PV4SCIE ← SciScore	0.949	0.938
PV5SCIE ← SciScore	0.952	0.94
PV3SCIE ← SciScore	0.952	0.941
STEM ← SciScore	0.173	0.151
PV2MATH ← MathScore	0.935	0.928
PV1MATH ← MathScore	0.934	0.924
ST67N01 ← SciScore	0.08	0.092
ST67N01 ← MathScore	0.074	0.087
ST67N01 ← SelfEfficacy	-0.019	-0.049
ST67N01 ← ST62N04	0.042	0.042
ST62N01 ← ST62N02	-0.268	-0.828
STEM ← ATTCOMP	0.048	0.018
HISCED ← IMMIG	0.123	1.04
IMMIG ← ATTCOMP	0.154	0.177
IMMIG ← HISCED	0.001	-0.053

ATTCOMP ← IMMIG	0.445	0.102

Table 7 SEM model standardized regression weights

The standardized regression coefficients are the values associated with each path. These numbers indicate the change in Y as a product of the standard deviation unit change in X. We can assess the relative magnitude of the effects of different explanatory variables in the model using standardized coefficients. From examining Table 7 above, for males, the highest regression weight is ST67N02 ← MotivationSci which reveals that for every increase of one standard deviation in the ST67N02 (which denotes whether a respondent wishes to study science in higher education), it increases their motivation in science by 0.976 of a standard deviation, nearly a 1:1 ratio. This is to be expected as a student who studied science in school, motivations to continue studying science in third-level education would be expected to be higher than someone who did not study science, this is mirrored in the female coefficient which stands at 0.984. Another expected result is that all five plausible maths values and all five plausible science values have a standardized regression weight of at least .926 with their respective subjects. Interestingly, out of the four observed variables used to create the latent variable SelfEfficacy (ST62N01, ST62N02, ST62N03, ST62N04), the highest regression weight for both genders was ST62N01 which represents how well a respondent felt they were performing in English. The female's regression coefficient is 1.301 which is the highest of any regression coefficient revealing an increase of 1.3 standard deviations in SelfEfficacy for every one of the variables ST62N01. The biggest difference between males and females is the regression coefficient between immigration status and attitude towards computers, with men at a significant value of 0.445 compared to women with only 0.102 this suggests that male non-natives have a higher interest in computing. As previously mentioned, having role models could potentially motivate an individual's decision to do STEM, this is supported by the SEM model with regression coefficients of 0.31 and 0.278.

Path	P-value
Full model	0
ST62N02 ← SelfEfficacy	0.037
ST62N03 ← SelfEfficacy	0.025
ST62N04 ← SelfEfficacy	0.014
ST67N03 ← MotivationSci	0.014
HIGHCONF ← SelfEfficacy	0
ST67N01 ← SciScore	0
STEM ← SelfEfficacy	0.010
HISCED ← IMMIG	0.037
ST62N01 ← ST62N02	0.035

Table 8 Multigroup Analysis Summary

As mentioned, the multigroup analysis is the result of a chi-squared difference test therefore P-Values below 0.05 are significant. The paths that are significantly different for each groups model are listed above in Table 8. The main thing to note from these results is that there is a clear difference in the overall models, but more specifically in three out of the four observed variables used to create the latent variable SelfEfficacy along with the effect this variable has on the outcome variable STEM. This suggests that self-efficacy is one of the biggest contrasts between the two groups.

	M1 Complete	M1 Imputed	M2 Complete	M2 Imputed
Chi-Square	2028.615	1547.548	2028.615	1981.254
P-value	0	0	0	0
GFI	0.9712132	0.9912152	0.9712132	0.9789142
AGFI	0.9546874	0.9855648	0.9546874	0.9668094
NFI	0.9828866	0.9952927	0.9828866	0.9876462
TLI	0.9815534	0.9951863	0.9815534	0.9867294
CFI	0.9846752	0.9965822	0.9846752	0.9889752
RMSEA	0.04013157	0.01948834	0.04013157	0.03433582

Table 9 SEM evaluation metrics R

The SEM model was also built in R using the Lavaan package. The results of the different models are shown in Table 9 below. In this case, all the model's evaluation metrics indicated good fitting models with the imputed data yielding better results.

5.3. Predictive Models

All the different model outcomes were stored in confusion matrices because they were predictive classification models. In the field of machine learning, a confusion matrix is a table structure that visualizes the performance of an algorithm, usually a supervised learning algorithm, in the task of statistical classification. We can find out how many true positives, true negatives, false positives, and false negatives there are in the confusion matrix. Accuracy, sensitivity, specificity, and Kappa are some of the performance criteria it provides. In classification problems, accuracy refers to the number of correct predictions made by the model across all types of predictions. When the target variable classes in the data are approximately balanced, accuracy is a good measure. Sensitivity informs us about a classifier's effectiveness in terms of false negatives; if we want to focus more on reducing False Negatives, we want sensitivity to be as close to 100% as possible. Specificity is the exact opposite of sensitivity, which is the proportion of genuine positives. For classification accuracy, Kappa or Cohen's Kappa is normalized at the baseline of random chance on your dataset. It's a better metric to utilize when there's an unequal distribution of classes. Because we have a strong class balance, the most significant evaluation matrix for this prediction would be accuracy.

5.3.1. Logistic Regression

The table below, Table 10, shows the results of the full model and the optimised Logistic Regression model completed with both the complete case data and the imputed data.

	M1 Complete	M1 Imputed	M2 Complete	M2 Imputed
Accuracy	0.7073	0.6959	0.7059	0.6977
Kappa	0.3676	0.3876	0.3646	0.3918
Sensitivity	0.5716	0.6682	0.5703	0.6769
Specificity	0.79	0.7193	0.7884	0.7153

Table 10 Logistic Regression evaluation metrics

The best Logistic Regression model is model one which is the full Logistic Regression model using all available variables to predict the response variable applied to the complete case data. This model has the highest accuracy as well specificity.

5.3.2. Naïve Bayes

The table below Table 11 shows the results of the full and the tuned Naïve Bayes model completed with both the complete case data and the imputed data.

	M1 Complete	M1 Imputed	M2 Complete	M2 Imputed
Accuracy	0.6922	0.6806	0.6245	0.5492
Kappa	0.3423	0.3468	0.0206	0.0053
Sensitivity	0.5825	0.6033	0.02911	0.01276
Specificity	0.7586	0.741	0.98764	0.9921

Table 11 Naïve Bayes evaluation metrics

The best naïve Bayes model is also the full model accomplished using the complete cases data, this model has the highest accuracy although the tuned models had much higher specificity, they had a really poor sensitivity as well as kappa therefore the full model is the superior choice.

5.3.3. Decision Tree

Table 12 below, shows the results of the Decision Tree model completed with C50, rpart and the bagged model. All models were completed using the complete case data and the imputed data.

	M1 Complete	M1 Imputed	M2 Complete	M2 Imputed	M3 Complete	M3 Imputed
Accuracy	0.6867	0.6778	0.6867	0.6804	0.6871	0.6818
Kappa	0.3266	0.3527	0.3266	0.351	0.3278	0.3542
Sensitivity	0.5577	0.6635	0.5577	0.5967	0.5589	0.6014
Specificity	0.7653	0.69	0.7653	0.7513	0.7653	0.75

Table 12 Decision Tree evaluation metrics

The best Decision Tree model is the bagged model also using the complete cases data, this had the best overall accuracy and specificity when compared to the rest of the models.

5.3.4. Random Forest

The table below, Table 13, shows the results of the full and the pruned Random Forest model completed with both the complete case data and the imputed data.

	M1 Complete	M1 Imputed	M2 Complete	M2 Imputed
Accuracy	0.6795	0.6894	0.6843	0.6825
Kappa	0.2829	0.3663	0.3066	0.3566
Sensitivity	0.4487	0.5755	0.5006	0.6116
Specificity	0.8201	0.786	0.7961	0.7427

Table 13 Random Forest evaluation metrics

Upon examining the results, we found that model one using the complete case data proved to have the highest specificity however model two using the imputed data had the best sensitivity. Despite this, model one using the imputed data had the highest accuracy and kappa and therefore was chosen as the optimal model. Interestingly, this is the only optimal model that consists of imputed data.

5.3.5. Comparison

	Logistic Regression	Naïve Bayes	Decision Tree	Random Forest
Accuracy	0.7073	0.6922	0.6871	0.6894
Kappa	0.3676	0.3423	0.3278	0.3663
Sensitivity	0.5716	0.5825	0.5589	0.5755
Specificity	0.79	0.7586	0.7653	0.786

Table 14 Optimal model comparison

Overall, for predicting if a respondent studies a STEM subject in school I found Logistic Regression using all variables to predict the outcome variable STEM applied to only

complete case data to be the most suitable method as it had the highest accuracy, kappa and specificity. Although it didn't have the highest sensitivity it was only 0.0109 away from the model with the highest sensitivity which was Naïve Bayes. A comparison of the different models can be seen in Fig. 53. Although Logistic Regression was the best model out of all the models we tested, none of the models are completely optimal. All kappa's are within the range of 0.21–0.40 which is interpretated as fair but it's not until over 0.40 that is moderate and 60 as substantial. In all cases a value of 1 is optimal, leaving us with a lot of space and need for improvement.

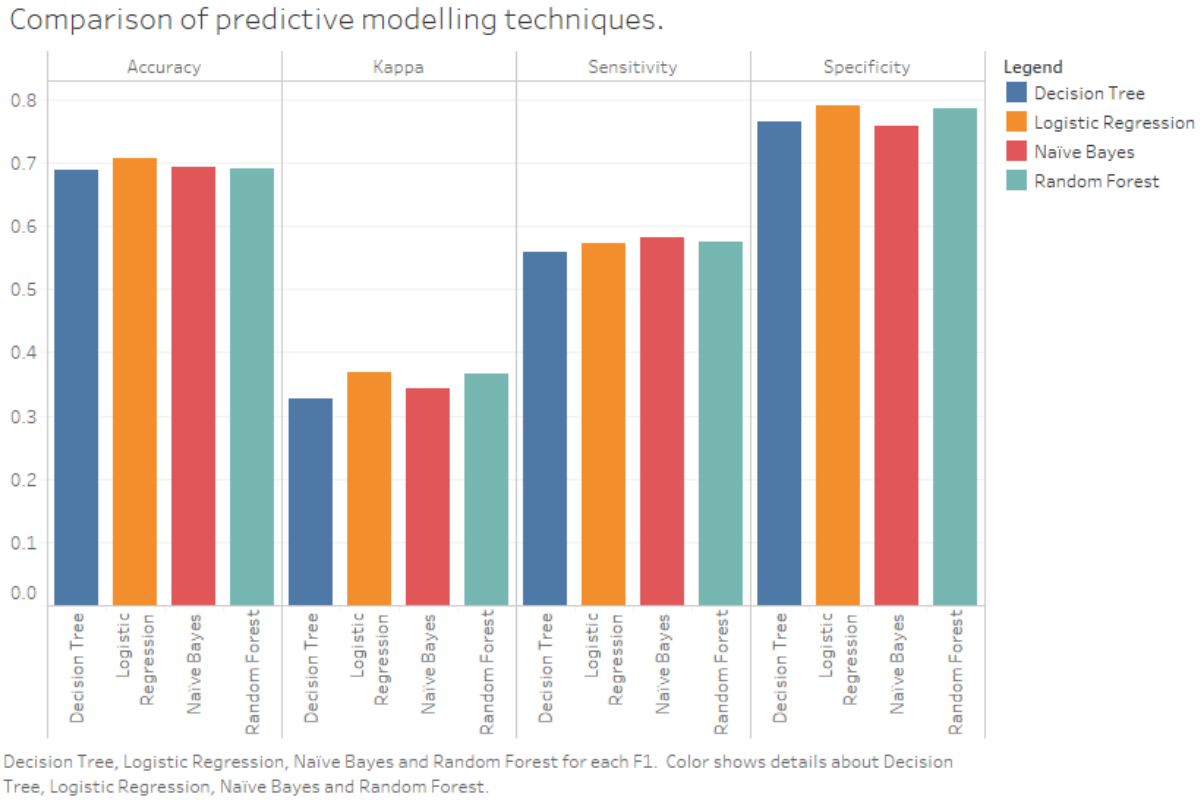


Fig. 53 Model comparison

6.0 Conclusions

The topic addressed in this report resulted in the construction of a data warehouse in the form of the SQL server; I detected relevant patterns by analysing my data, and I inquired about the gender gap in STEM fields. I also created the SEM model I wanted for the inquiry, examined it, and came to the conclusion that a predictive model was possible. Finally, I compiled all of my findings and reconciled my findings so that I could make recommendations for the issues I had identified. The objectives I listed indicate the success criteria or definition of done that was defined in the CRISP-DM model's initial stage, and this is the criteria by which I declare the project finished.

I worked on the back of confidence that I had gained from testing and tuning the models which assured me I would be generating the best possible results that I could. I took these extra steps to ensure that I could get the best results: I tested all models with data that contained only complete cases and data that contained imputed data to ensure that the correct method for dealing with missing data was chosen; I tested models with different train and test splits to find the split that would provide the best class balance. I used seeds to ensure that whatever results I got were repeatable. I created several different models in order to determine which ones were the most effective. I ran tests to check that the data's assumptions were not violated. One noteworthy limitation which the project suffered was a lack of broad data, because the dataset was domain-specific to Australia, a wider-reaching dataset would have offered a geographical dynamic to the investigation and allowed for country-by-country comparisons.

From the study, I can conclude that difficulty, a lack of support, a lack of role models within the field, a sense of not belonging, prejudice, a lack of self-efficacy, and true disinterest are all issues and proven deterrents that keep women out of STEM. Although it has been demonstrated to be an influence, the majority of STEM workers state they did not find the topics simple in school. This could be due to bad teaching methods; possibly making the content more approachable or practical could draw more students to the classes. Female-oriented bootcamps, coding events, or projects might help minimize the feeling that they don't belong in a STEM classroom due to a lack of role models both professionally and personally. Another possible cause is that the minority of women in some classrooms can be off-putting to some. Many students may feel under-supported at home because their parents or guardians are ill-equipped to solve technical or advanced syllabuses and thus are unable to assist their children as much as they would like to. Additional supports for both parents and students can help add additional resources to enhance the learning experience.

While a lack of interest will always be a factor in why some students choose not to pursue STEM, the only effective approach to reduce this is to expose children to STEM in a way that allows them to make an informed decision about whether STEM is suitable for them. Self-efficacy is a key barrier to females entering STEM fields; both science and math heavily influence a student's self-efficacy; although women appear to link their self-efficacy more to their English scores, in general, academic achievement appears to be a decisive factor in a student's confidence. Teaching malpractice, bullying, and extremely competitive and degrading class situations can all contribute to low self-efficacy. In order for pupils to thrive, they must be encouraged by their failures and motivated to succeed. Educators of all kinds must take care to build rather than tear down students' confidence. Discrimination is also a major problem that is tough to fix and has been going on for a long time. Raising awareness and educating people about misogyny in the workplace is a good place to start. In order to achieve a balanced and proportionate representation of women in STEM fields, we still have a long way to go.

All visualizations can be found in my interactive tableau dashboards found at the following links:

https://public.tableau.com/app/profile/zara.o.brien/viz/FYP_16273382656330/Dashboard1

https://public.tableau.com/app/profile/zara.o.brien/viz/WomeninSTEM2/Dashboard2

https://public.tableau.com/app/profile/zara.o.brien/viz/WomeninSTEM3/Dashboard3

https://public.tableau.com/app/profile/zara.o.brien/viz/WomeninSTEM4/Dashboard4

7.0 Further Development or Research

My research was confined to Australia because it was the only country that provided the necessary data for the investigation; however, if more data becomes available, I would like to compare different countries and areas to add an intriguing dimension to the analysis.

Although I was able to anticipate the uptake of degrees for the following year, I would like to make a larger forecast with more time-based data so that I could look further into the future to assess the direction of the current trend.

Even though I was able to design my tableau dashboard, I would have preferred more time to devote to it in order to improve and develop it, since I did not believe I had enough time to bring it up to the standard I wanted to present it at.

8.0 References

Australian Government (2021) 'STEM fields of education and research'. Department of Industry, Science, Energy and Resources. Available at: https://www.industry.gov.au/data-and-publications/stem-equity-monitor/methodology (Accessed: 27 July 2021).

Australian Government Department Of Education, S. (2020) 'Longitudinal Surveys of Australian Youth, 2009 cohort (Version 9.0)'. ADA Dataverse. doi: 10.4225/87/6BW27V.

Bentler, P. (2001) 'Structural Equation Modeling', *Structural Equation Modeling*, p. 8.

Brown, T. A. (2006) *Confirmatory Factor Analysis for Applied Research*. Available at: https://books.google.ie/books/about/Confirmatory_Factor_Analysis_for_Applied.html?id=KZwDkH2G2PMC (Accessed: 28 July 2021).

Choney, S. (2018) 'Why do girls lose interest in STEM? New research has some answers — and what we can do about it', *Stories*. Available at: https://news.microsoft.com/features/why-do-girls-lose-interest-in-stem-new-research-has-some-answers-and-what-we-can-do-about-it/ (Accessed: 26 July 2021).

Correll, S. (2001) *Gender and the career choice process: the role of biased self-assessments | Sociology*. Available at: https://sociology.stanford.edu/publications/gender-and-career-choice-process-role-biased-self-assessments (Accessed: 26 July 2021).

DHS (2016) 'STEM Designated Degree Program List'. Available at: https://www.ice.gov/sites/default/files/documents/Document/2016/stem-list.pdf?fbclid=IwAR1B6RLA8ITdt5PQ8V_ZW-cWms_42Ppgio699ZAHTU2cngF7Xpm12fxFOsA.

Ellis, J., Fosdick, B. K. and Rasmussen, C. (2016) 'Women 1.5 Times More Likely to Leave STEM Pipeline after Calculus Compared to Men: Lack of Mathematical Confidence a Potential Culprit', *PLOS ONE*, 11(7), p. e0157447. doi: 10.1371/journal.pone.0157447.

Funk, C. and Parker, K. (2018) 'Women and Men in STEM Often at Odds Over Workplace Equity', *Pew Research Center's Social & Demographic Trends Project*, 9 January. Available at: https://www.pewresearch.org/social-trends/2018/01/09/women-and-men-in-stem-often-at-odds-over-workplace-equity/ (Accessed: 9 April 2021).

Hooper, D., Coughlan, J. and Mullen, M. (2007) 'Structural Equation Modeling: Guidelines for Determining Model Fit', *The Electronic Journal of Business Research Methods*, 6.

Jeffries, D., Curtis, D. D. and Conner, L. N. (2020) 'Student Factors Influencing STEM Subject Choice in Year 12: a Structural Equation Model Using PISA/LSAY Data', *International Journal of Science and Mathematics Education*, 18(3), pp. 441–461. doi: 10.1007/s10763-019-09972-5.

Lazio, R. and Ford, H. (2019) *The U.S. Needs to Prepare Workers for STEM Jobs*, *SHRM*. Available at: https://www.shrm.org/hr-today/news/hr-magazine/summer2019/pages/the-u.s.-needs-to-prepare-workers-for-stem-jobs.aspx (Accessed: 26 July 2021).

Loewus, L. (2015) 'When Did Science Education Become STEM?', *Education Week*, 2 April. Available at: https://www.edweek.org/teaching-learning/when-did-science-education-become-stem/2015/04 (Accessed: 26 July 2021).

National Center for Education Statistics (2019) 'Number and percentage distribution of science, technology, engineering, and mathematics (STEM) degrees/certificates conferred by postsecondary institutions, by race/ethnicity, level of degree/certificate, and sex of student: 2008-09 through 2017-18'. National Center for Education Statistics. Available at: https://nces.ed.gov/programs/digest/d19/tables/dt19_318.45.asp?current=yes (Accessed: 27 July 2021).

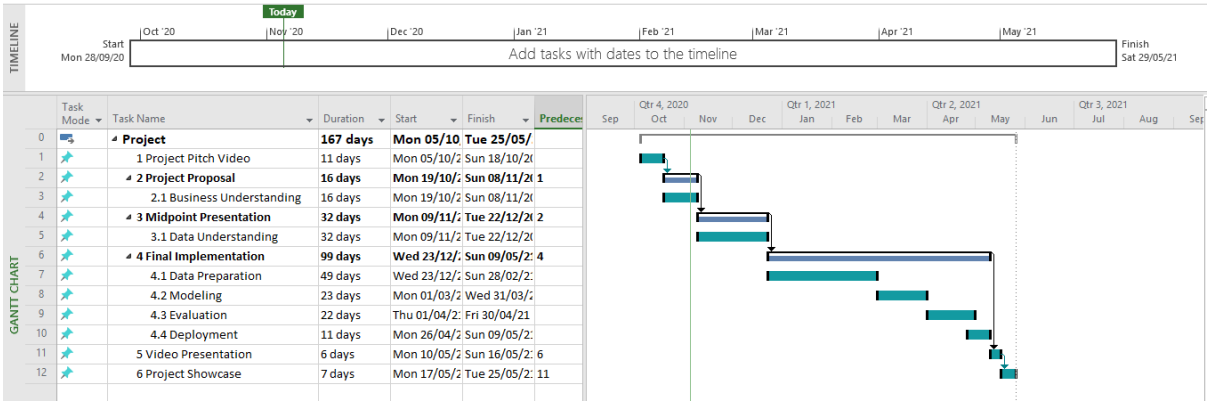
O'Callaghan, C. (2021) 'What is STEM?', *Top Universities*. Available at: https://www.topuniversities.com/courses/engineering/what-stem (Accessed: 26 July 2021).

Pew Research Center (2017) 'STEM Survey'. Pew Research Center. Available at: https://www.pewresearch.org/social-trends/dataset/2017-pew-research-center-stem-survey/ (Accessed: 27 July 2021).

Schermelleh-Engel, K., Moosbrugger, H. and Müller, H. (2003) 'Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures', *Methods of Psychological Research Online*, 8, pp. 23–74.

9.0 Appendices

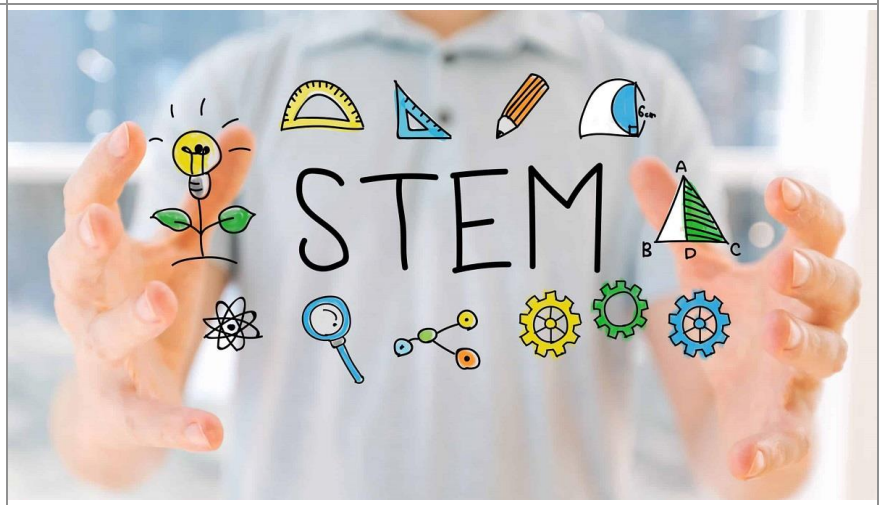
9.1. Project Plan



9.2. Showcase Profile

Student Name:	Zara O'Brien
Programme:	BSc (Honours) in Computing
Specialisation:	Data Analytics
Personal Bio:	Completed a six month work placement as an IT support intern at Arthur Cox, gaining experience in setting up laptops for users, working on a IT helpdesk, setting up remote access for users and assisting with inventory management. Strong skills in R, SPSS and data analysis. Accumulated skills in Java, SQL, Python and HTML. Strong team player and communication skills developed during college projects. Good organisation and planning skills developed with the experience of organising fundraising events. Seeking data analytics opportunities on graduation.
Project Title:	Analysis of women in STEM.

Project Overview:	The STEM industry is one of the biggest global sectors having grown by 79% since 1990. The Covid-19 pandemic has highlighted the necessity of technology in our lives as well as the need to be proficient with technology. Although the sector is widespread, women make up only 18% of computing undergraduates and 28% of the science and engineering workforce in the US. As a woman in STEM I have seen the gender inequality both in college and in the workforce, giving me an interest in the topic. The research conducted would also be of particular interest to educators and professionals in the STEM industry. This project investigates the polarisation in the number of women to men in STEM (Science, Technology, Engineering and Mathematics) subjects. After establishing a divide, the underlying factors were individually examined to measure their effect on an individual's decision to pursue STEM. Multiple machine learning models were then developed attempting to predict if a student would study a STEM subject. The results were then compared, to examine which of the models was the most optimal. The longitudinal surveys of Australian youth data was used, this was authorised by the Australian Data Archive. This data combines PISA (Programme for International Student Assessment) data along with survey responses for the following decade. This dataset was initially examined in SPSS before being stored in Microsoft SQL Server where it could be accessed directly in Alteryx, Tableau and R Studio to be analysed.
Technologies Used:	R, R Studio, SPSS, Tableau, Microsoft SQL Server, Excel, Alteryx
LinkedIn Profile URL:	https://www.linkedin.com/in/zaraobrien/
Link to Portfolio (GitHub Repo):	

Profile Photo:	
Project Image 1:	
Project Image 2:	

Project Image 3:	

9.3. Showcase Poster

National College of Ireland

Analysis of Women in STEM

Zara O'Brien - BSc Computing
(Data Analytics)

Despite the fact that STEM (Science, Technology, Engineering, and Math) employment has grown by 79% since 1990, women make up only 27% of the STEM workforce as of 2019. Understanding the factors that contribute to the underrepresentation of women in STEM areas is critical if more women are to be encouraged to pursue careers in the subject.

- The data for this investigation was acquired from National Center for Education Statistics (NCES), Pew Research Center and the Longitudinal Surveys of Australian Youth (LSAY).
- The data was analyzed using a variety of techniques in order to infer intriguing findings that were utilized to steer the overall project.
- Structural Equation Modelling (SEM) was used to gain insights into a student's decision to pursue STEM and the factors that influence it.
- Multiple Logistic Regression, Naïve Bayes, Decision Tree and Random Forest algorithms were used to try to predict whether a student would elect to study STEM during their final year of school.

Data

Analysis

Modelling

Microsoft SQL Server

tableau

RStudio

SPSS AMOS 26

9.4. Reflective Journals

9.4.1. October

Student name: Zara O'Brien

Student number: X17363043

Programme: BSHC (Hons) in Computing

Specialization: Data Analytics

Month: October 2020

This is my first reflective journal for my final year project. Over the seven months of my project, I will take this time to reflect, analyse, and try to work out the best way to move forward. I think this will be extremely useful to see how my project is progressing, what I need to work on, and how well I am managing my time and sticking to my project timeline. Each of my monthly journals will follow Driscoll's (2000) Model of Reflection as recommended by my lecturer in my project class.

WHAT?

> This month mainly involved me coming up with my project idea and then pitching that idea to my lecturers. I had brainstormed some ideas over the summer, once I was informed that I had got the specialization of data analytics I started to think about my project and what I could do for it. Our project class was the first class of the year, and it was explained to us just how important the project would be as it is what we will take with us to show future employers. The first task was to choose our idea, at this point I had already made a list of ten ideas that I found interesting. I decided to take three of these and look at them more in-depth to choose what I felt would be the best project. I decided that I would like to do an analysis of women in STEM. Once I had my idea the next step was to make my project pitch video this covered what data would be used for the project, where this data would be sourced, the aims of the project, why it is challenging, who the project is for and how it is different to what had been done before. I uploaded this video to be reviewed.

SO WHAT?

> This month was perhaps most important in terms of my project as without my idea I have nothing to work on. I was nervous doing my project pitch video, but it was necessary to express my idea. The video was to be no longer than five minutes in length, initially, I had thought I wouldn't be able to stretch it out but when I recorded it the time went by really quick. Since uploading my project pitch video I am slightly anxious that my idea isn't good enough or possibly not complex enough, however, I am welcome to any changes and improvements that could be suggested. I am now waiting to get feedback on my idea and will use this to do my project proposal.

NOW WHAT?

The next step in my project is to see if my project idea has been accepted and then to meet with my project supervisor to talk through and finalise the idea. I will then finish my project proposal and make a project plan. Part of the project proposal is making a project plan, this will be extremely important to help with time management and seeing where I stand at any given point in the project timeline. I aim to spend a good amount of time planning out my project carefully as I think this will set the pace for my project and will make sure that I get everything completed in time. Once my project proposal is submitted along with my ethics forms I will work towards my mid-point presentation. I am still looking for data to use in my project as I am worried I won't have enough data to complete a significant analysis.

9.4.2. November

Student name: Zara O'Brien

Student number: X17363043

Programme: BSHC (Hons) in Computing

Specialization: Data Analytics

Month: November 2020

WHAT?

This month my project idea was accepted and I then wrote my project proposal. The project proposal involved writing about different aspects of my project that were to be considered. This included background, technical approach, special resources required, project plan, technical details and evaluation. I have been continuing the search for data and now have nineteen different sources. Along with the proposal, I uploaded my ethics form. For the ethics form, I gathered permission to use each dataset either from the website or via email. Since uploading the project proposal, I have been working towards the midpoint presentation due in December. The project proposal I have already uploaded is not the final upload, however, the final upload is required with the midpoint presentation along with my preliminary analysis and evidence of technology used. Similarly to most students in fourth year, this month had been extremely busy with a lot of deadlines meaning I could not spend as much time as I would have liked to on the project.

SO WHAT?

Doing my project proposal was a big step in the planning and preparation of my project. By doing the proposal and ethics form I was forced to consider all the aspects of my project which needed to be done. On paper, it was easy to describe out how I would do my project and what tools I would use but in practice, things don't always go to plan and some things are overlooked. Since I started to look at and work towards the midpoint presentation I have realised that the technologies that I laid out in my project proposal may not be the best fit for the project. I am now slightly more concerned with the architecture of my project and think this needs to be carefully thought through. Although most of my time was used on other modules some of them were relevant to the project. I completed a data application development CA which involved selecting, pre-processing, transforming, mining and interpreting data as well as following a methodology and writing a report. I believe this will aid me in this project.

NOW WHAT?

Now I need to focus on revising my project proposal and making sure that my project will be viable. The midpoint presentation is dependent on this so they will work hand in hand. For the midpoint presentation, I need to demonstrate how my project will

work. By figuring this out and demonstrating the storing, retrieval and manipulation of my data I will be able to amend my project proposal to reflect this and know exactly what tools and technologies to use.

9.4.3. December

Student name: Zara O'Brien

Student number: X17363043

Programme: BSHC (Hons) in Computing

Specialization: Data Analytics

Month: December 2020

WHAT?

This month has seen a big shift in the focus of my project. I needed to focus on something more specific and this took a lot of time and consideration. Eventually, I decided to take a closer look at why fewer girls pursue STEM particularly if self-efficacy played an important role. Once my focus changed there was a lot of work to do before the midpoint presentation. My first step was to change my project proposal to fit in with my new focus and from that get an overview of what tools and technologies I would need to use for my mid-point. I then needed to pick specific data sets, clean and prepare them and conduct a preliminary analysis on them to gain some insights. By completing my midpoint presentation I tied down my project idea, decided on what data and technologies to use, conducted a preliminary analysis and I am now happy with how my project is progressing.

SO WHAT?

The last month in college has certainly been the most stressful for all modules with the software project being no exception. By needing to rethink the focus of my project I felt very behind and overwhelmed and this was hard to manage at the time. Once I knew what I wanted to do I was still very busy and it was stressful but it was just a matter of putting a lot of work in. Although it was a struggle to complete the mid-point presentation on time I feel in a much better position moving forward with my project and I am more confident in my idea. If this shift in focus had not occurred I think that my project would have started to fall apart later in the year which would have resulted in it being too late to change anything and ultimately having very little to work with. So although it was difficult at the time I think it was necessary for the future of the project. I would imagine that this is a different experience to most other people who were happy with their idea, however, it seemed to be a very demanding month for everyone.

NOW WHAT?

Now I will wait to get feedback on my midpoint to get a better idea of how my project is going and anything I can do to improve it. I will continue working on the analysis of the data to get concrete figures to use in my project. As I don't have much experience using Tableau I will try to get used to the software in my spare time so that I will be very familiar with it when it comes to making my final visualizations.

This semester I will also start to learn about machine learning and making predictive models which I will need to know for my project. This will involve me skipping ahead of my modules as to not leave it to the end of my project to get started on this section.

9.4.4. January

Student name: Zara O'Brien

Student number: X17363043

Programme: BSHC (Hons) in Computing

Specialization: Data Analytics

Month: January 2021

WHAT?

Since uploading my mid-point I took some time off as the last few weeks of semester one was extremely demanding. I have had one week of my new modules which I am excited about as they seem really interesting and beneficial for my final year project. While waiting for my results and feedback from my mid-point I set up Microsoft MySQL and connected this with tableau as I knew I would need to store my data and that I wanted to use tableau for my visualizations. After receiving my results and feedback I know what I need to focus on going forward with my project.

SO WHAT?

This month did not see as much progress as I would have liked, however, as I changed my project quite a bit for the mid-point I wanted to wait and get feedback from that. As I had made some big changes I was nervous about this feedback but I was happy with my result. The feedback from my mid-point was that I needed to fix some errors with the documentation of my project which I was expecting as I ran out of time to do this properly and to change the title of my project. My supervisor would like me to incorporate data engineering into my project and to identify what machine learning algorithms I will apply to the dataset. Overall the complexity of my project needs to be improved and I am not entirely sure how to do that so I am concerned about that at the moment.

NOW WHAT?

Now I really need to figure out how to increase my projects difficulty to improve my mark for the final submission. This will require a lot of research to find things that will fit well into my project. Once I find what I would like to add to make the project more difficult I will work towards incorporating that into my project.

9.4.5. February

Student name: Zara O'Brien

Student number: X17363043

Programme: BSHC (Hons) in Computing

Specialization: Data Analytics

Month: February 2021

WHAT?

This month saw another change in my project. To successfully apply predictive models to my data I needed a variable to use as the result, in this case, if they went on to study or work in STEM. I found and gained access to an academic paper that looks at factors that influence children choosing a STEM subject for their final year in school using a Structural Equation Model. This research used the PISA data that I have been using along with LSAY (longitudinal surveys of Australian youth) data. I then had to apply for this data which was granted to me after a few weeks.

SO WHAT?

Once I received this data I had to take a step back and get back up to speed in terms of having the data cleaned and stored ready to be analysed. The data I was using grew from 300 variables to over 7000 and preparing this new data took more time than I would have liked but was necessary to do. The data was much larger and more complex than just the PISA data I was using before. The LSAY data had many different formats and its answers were coded differently. The process of getting the data processed took longer than I would have liked but I am now happy to have it ready to work on again.

NOW WHAT?

I feel as though I have had multiple things slowing me down recently in the project so now I really need to get caught up and work hard to put myself in a better position next month. The next step I will take is to completely analyse the data and to apply different predictive models and compare their performances. Once I have this done I would like to move on to optimisation, testing and visualizations.

9.4.6. March

Student name: Zara O'Brien

Student number: X17363043

Programme: BSHC (Hons) in Computing

Specialization: Data Analytics

Month: March 2021

WHAT?

This month was tough in terms of my project, I thought I would be able to start applying models right away but I actually had a lot more preparation to do. As the data I am using is survey data the data that is missing is missing not at random, this means imputation would not be an appropriate method of dealing with the null values. Instead of replacing the values, I re-coded all my variables to factors with an NA level. To make this easier on myself I added functions to my Alteryx workflow to replace any values marked as NA, invalid or missing to NA. This sped up the process of re-coding as there are about seventeen different ways these values were coded into the original dataset which could all be put under the NA level when coded as factors. Something else I noticed was the number of variables I had even after I had taken out anything irrelevant along with some variables that gave no purpose alone. To solve this I did some feature engineering for example occupation codes alone served no significance to my analysis but when coded to filter down to just STEM occupational codes it was a lot more useful. Once this process was completed I began applying models to my data.

SO WHAT?

Unfortunately, the preparation for the modelling stage of my project took longer than expected however I have now started this stage of my project and am happy with the progress. Although it was a long process it taught me about how important data preparation for modelling is, something I never expected. Due to the last month not being as productive as I would have liked I will have to dedicate most of my reading week to getting back on track with the project which I am prepared to do.

NOW WHAT?

As stated above I will be spending most of my reading week working on the modelling stage of my project in the hopes to have this completed by the end of the week allowing me to move on to optimisation, testing and visualizations.

9.4.7. April

Student name: Zara O'Brien

Student number: X17363043

Programme: BSHC (Hons) in Computing

Specialization: Data Analytics

Month: April 2021

WHAT?

Although I had planned on catching up on project work over the reading week CA's from other modules needed my focus as they were due much sooner, leaving me struggling with the project. I ran into some trouble as I realised I had prepared my data incorrect for the model and perhaps had taken on too many different variables. At this point, I decided to reach out to one of last years students to get their advice on what to do as they were in a similar position with their project at this stage last year. They advised that I started over and focused on fewer variables to not overwhelm myself or the model.

SO WHAT?

Now trying to juggle CA's and restarting my project with one month to go I began to panic if I would get the project done and if I did it would not reflect my capabilities. At this stage, I decided to apply for a project deferral to give me a chance to submit a project I would be happy with.

NOW WHAT?

Now I have applied for a project deferral and I will focus on finishing my CA's until I hear back from the exams office to know where I stand with the deadline for the project.

9.4.8. May

Student name: Zara O'Brien

Student number: X17363043

Programme: BSHC (Hons) in Computing

Specialization: Data Analytics

Month: May 2021

WHAT?

This month I heard back from the exams office and got the confirmation that I was granted a deferral for my project. At the same time exams were coming up and I wanted to give them my all so I focused on them until they were due. After exams, I was really burnt out and sick so I needed to take two weeks off to recover, however during this time I made a new plan for the project going forward and tried to read up and watch videos on SEM models.

SO WHAT?

When making my new plan I decided I would focus on creating a SEM model to examine the decision-making process of students picking a STEM subject in their final year of school. Although this had been done in the paper I had previously found I was now able to access the data they suggested in their further work section which was not available to them allowing me to examine many more factors.

NOW WHAT?

Now I need to get stuck into my project following the plan I created to ensure I have enough time to complete the project to a high standard by August.

9.4.9. June

Student name: Zara O'Brien

Student number: X17363043

Programme: BSHC (Hons) in Computing

Specialization: Data Analytics

Month: June 2021

WHAT?

This month I put a lot of work into my project to really get it kick-started. Using the same data as before I set up SQL Server and uploaded my data here. I then started on my data preparation in both R studio and Alteryx, this included removing any noise or variables that I was not interested in, feature engineering to create some variables and changing scale variables so they were all formatted the same. I then went on to check the assumptions of the data using statistical tests in SPSS, here I tested for things such as multivariate normality, sufficiently large sample size, outliers and missing data.

SO WHAT?

The process of cleaning and preparing my data took a long time but as I learned previously it was a critical step in ensuring my project would work. This took a lot of time and effort but I am happy to put the time and effort into the project over the next month.

NOW WHAT?

Since getting the rest of my results at the end of this month it has given me a huge push to really work hard on my project. The next step in my project is to move onto the modelling stage and test the model accordingly before moving onto visualizations, giving myself some extra time for any issues that might arise.

9.4.10.July

Student name: Zara O'Brien

Student number: X17363043

Programme: BSHC (Hons) in Computing

Specialization: Data Analytics

Month: July 2021

WHAT?

This month was extremely busy for me with the project, I put in a minimum of eight hours a day to make sure I could get everything done in time and submit a project that I was happy with. I started with the modelling stage of my project which I ran into some issues with. The SEM model was a new concept to me so required a lot of research to gain an understanding of what exactly everything meant to ensure I built the model correctly, I tested the model as I went and once I had completed the SEM model I moved on to the predictive modelling. As I was more familiar with the predive modelling this section moved along quicker than the precious SEM model. I completed four different types of predictive models; Logistic Regression, Naïve Bayes, Decision Trees, and Random Forest. Each model was tested and then evaluated against one another to find the optimal model for the data used. I filled in the project document as I completed each section to ensure I did not leave too much till near the end. I then completed my visualizations in Tableau.

SO WHAT?

This month was stressful trying to get everything done in time for the submission but I was happy to put the work in to get it done. I probably spent too much time working on the modelling stage and did not leave enough time for the visualizations, although I was able to get them done in time I wasn't entirely happy with them and would have liked more time to improve on them.

NOW WHAT?

Now I have finished my project I feel very proud of what I have achieved and I feel as though I learned so much throughout the entire project experience.

National College of Ireland

Project Proposal

An analysis of the gender gap in STEM

05/11/2020

BSc (Hons) in Computing

Data Analytics

2020/2021

Zara O'Brien

x17363043

x17363043@student.ncirl.ie

Contents

Objectives

Objective 1: My first objective is to find the data that I will need to conduct my analysis. From an initial search, it seems inevitable that this will mean merging many datasets to cover all the information I will need to get the desired results. I also aim to create my own data set. I will need to ensure that I have permission to use any data that I would like to use in my project.

Objective 2: My second objective is to clean and prepare the data for use. This will involve removing irrelevant data, handling missing data, checking for duplicate data and ensuring the data is in the correct format. I will also need to combine different datasets.

Objective 3: My third objective is to establish that there is a gender gap based on the data. For this, I also want to explore if the gender gap differs across different disciplines within STEM subjects and how the gender gap differs in different countries.

Objective 4: My fourth objective is to try and explore why there is a gender gap? I will be looking at things such as mathematic ability and stereotypes, possibly female STEM roles in media.

Objective 5: My fifth objective is to predict if the gender gap will close and if so how long it will take based on past trends.

Objective 6: My sixth objective is to document and display my findings in a clear but interesting way as well as reaching some conclusions based on my analysis.

Background

The idea for my project came as a result of my personal experiences. I first gained an interest in technology in my transition year during secondary school when Dell came in to speak to us about careers in IT. After speaking to us they asked anyone who had an interest in a future career in IT to raise their hand and to my surprise, I was the only girl with my hand up. I had thought that college would be different but on my first day, I was surprised at the ratio of women to men, I think there were only about five women in my course in the first year. Last year I got to do a six-month work placement at Arthur Cox, during my time here I continued to notice this trend with only two out of the fourteen people in the IT team being women. As I'm in my final year I am thinking about work after college and wondering if this trend will persist? has it always been this way? and why is it like this?

From researching the topic I found that these are questions that have been asked before and it became very obvious that there is a significant gender gap in STEM (Science,

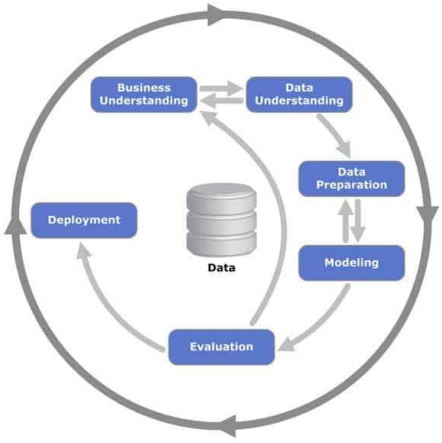
Technology, Engineering and Mathematics). This is a topic of much discussion as companies want to try to support minorities to encourage them to pursue careers in areas where they may be outnumbered. From looking at studies it has been proven that both female and male students have similar levels of skill and participation in math and science, so why is nursing so female-dominated? and why is engineering male-dominated? I read many articles about why there are overall fewer women studying and working in STEM than men and each gave their reasons - gender stereotyping, the wage gap, lack of family-friendliness, lack of female leaders and role models, low Confidence, workplace Bias etc...

From this research, I established that there are many articles on the topic of gender and STEM and they are supported by some statistics. There is also much discussion on why the gap exists however, there is not much data and what is there hasn't been brought together to create a bigger picture. So far I have only encountered one occurrence of someone making a prediction, this was based on academic publishing and left out engineering. For this reason, I think that my project will be different from what has been done before and will give a better overall analysis.

Technical Approach

CRISP-DM:

For my project, I will follow the CRISP-DM (CRoss-Industry Standard Process for Data Mining) approach as I think it will help to plan, organize, and implement my project and it has been the most widely used methodology since 2002. I will describe the six stages of this methodology below.



Business understanding: This involves making a preliminary plan for the project and understanding the aims/objectives of the project.

Data understanding: This step involves the collection of all the data required for the project and looking at this data to get first insights into what the data could show.

Data preparation: This is the process of constructing the final dataset to be used. This will involve the creation of a database and the cleaning of data.

Modelling: This stage is where any modelling techniques are applied to the dataset.

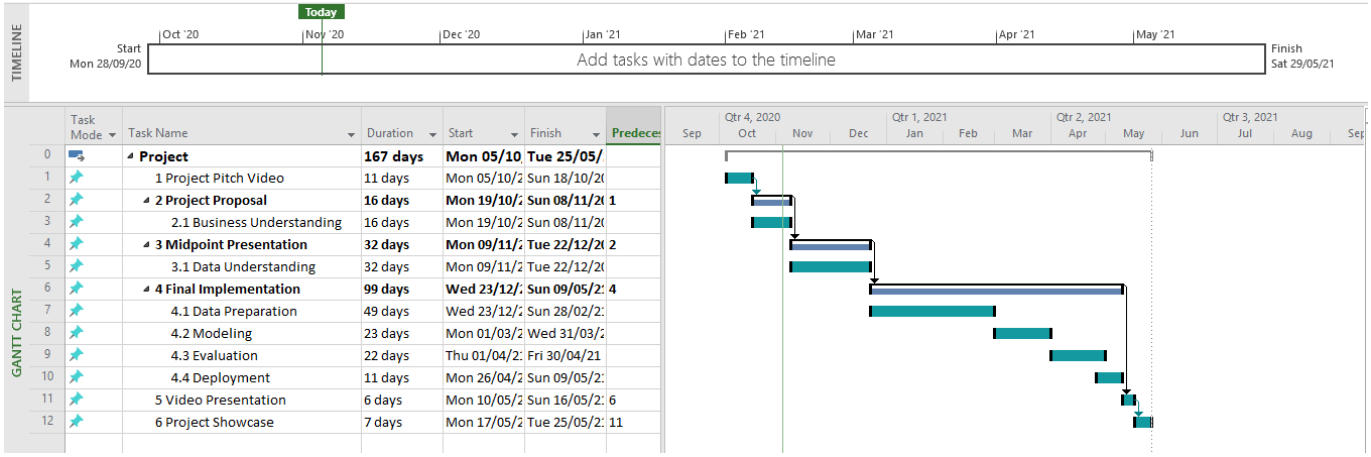
Evaluation: In the evaluation stage I will check to ensure that my project has properly achieved the objectives I will review the work to see if anything needs to be changed.

Deployment: Once I am happy with the results this is where I will actually create the model and display the results of the project.

Special Resources Required

Currently, I require no special resources.

Project Plan



Technical Details

R Language: R is a programming language ..” for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.” I will be using R to write my project and to display results.

R Studio: I will be using R Studio to build my project. R studio is an open-source integrated development environment for R Language. I have not yet chosen the libraries I will import into R studio to use for my project.

Microsoft SQL Server: Microsoft SQL Server is a relational database management system that stores and retrieves data to other software, so this will be perfect for holding my data. I will use this to store my data.

Tableau: Tableau is a free data visualization software. I will use tableau to display all of my findings and my final figures. Tableau dashboards are visually appealing and user friendly so I think it is the best option for displaying my results.

SPSS: “SPSS Statistics is a software package used for interactive, or batched, statistical analysis” and is one of the most popular statistical tools. I will mainly be using SPSS for testing, just to ensure accuracy and to get quick ideas about how things will turn out.

Excel: Excel is probably the most popular spreadsheet application with built-in statistical and graphing commands. I will be using Excel for my initial look at the data and to look at during the data exploration stage of my project.

Evaluation

I will evaluate my project throughout the different stages making sure the data I use fits in with the project and is the correct data to be using. I will also test my data accuracy by checking my calculations in different environments. I will conduct a model evaluation to make a forecast for known data to see how accurate the model is. I will make sure my objectives are met and that the findings are clearly displayed to end-users. To make sure the data is easily interpreted I will ask someone to look at my findings and check that it makes sense to them.