# Introduction

## Objectives

The brief for this project was as follows:

1. Conduct a survey with Maynooth University postgraduates.
2. First, decide on a topic of interest for the survey.
3. Then design the survey questions and carry out the survey.
4. Finally, analyse and interpret the results.

## Approach (Methods)

We decided to respond to the above project brief as follows:

1. Target the survey on our peers in the following courses:
   1. Higher Diploma in Applied Computer Science;
   2. Masters in Data Science;
   3. Higher Diploma in Data Science;
   4. Masters in Software Development.
2. Focus on the target population’s attitude to workload demands of their respective courses;
   * In particular, we were interested in establishing if a link existed between the reported overall **workload satisfaction** levels and **lecture attendance** rates.
3. Refer to Appendix A for a copy of the questionnaire that we designed for this survey
   * In this report we provide an overview of the data that was collected using that questionnaire during the week commencing Monday, 1st April 2019.
4. Use **SAS** (refer to Appendix B for the resultant code) to analyse the quality of the collected data – both empirically and graphically (using as wide a variety of graphical formats as we could devise) and to use suitable imputation approaches to address any data quality issues detected. (See the **Data Analysis** and **Data Imputation** sections below for further details.)

* We use both **Logistic Regression** and **Random Forest** learning techniques to explore for patterns in the collected data. (See the **Data Modelling** section below for further details.)

## Findings (Results) & Lessons Learnt (Discussion)

We found that, surprisingly, we could detect **no relationship** between **lecture attendance** and the level of student **satisfaction with course workload**. Instead, a combination of **Age, Level, Course, Gender and Job** were determined to be the most powerful predictors.

We learned that paper-based surveys are difficult to conduct and ensure good data quality during collection and data entry.

# Data Collection (Design & Execution)

Ideally, we would have liked to use a tool like **Survey Monkey** to conduct our opinion poll electronically but were unable to get the e-mail addresses we needed from the university administration – for, understandably, **data privacy** reasons. Therefore, taking a **paper-based approach** was our only alternative.

Having selected the focus area for our survey, our first task as a team was to decide what questions to pose and to design the **questionnaire** that we would use to conduct the survey. **Refer to Appendix A** for the resultant work product. We then tested this questionnaire by completing the survey ourselves and them making some (minor) modifications that we agreed.

All team members then set about using printed questionnaires to survey the students of the 4 postgraduate courses we had selected during the **week commencing Monday, 1st April 2019**. By the end of that week, we had collected 54 completed questionnaires. While this does not appear to be a large number in itself, it did provide us with a significant sample of the target student population.

Finally, we used **Microsoft Excel** to convert the contents of those completed questionnaires to electronic form (in CSV file format). Note that, while we did capture the **Student Numbers** of the respondents, we decided to replace them with automatically generated unique identifiers to ensure **anonymity**.

# Data Analysis (Empirical & Graphical)

After importing the resultant CSV file in to SAS we used a combination of **CONTENTS, MEANS, FREQ** and **SQL** procedure calls to analyse both the dataset as a whole and variable by variable. We also produced an array of graphs of various types to visualise key aspects of that analysis using a combination of **SGPLOT, SGPANEL, TRANSPOSE** and **SQL** procedure calls.

The following graphs illustrate just a **sample of the different visualisation formats** that we used for this project. (See the **SAS code in Appendix B** for further details.)

|  |  |
| --- | --- |
| Figure . Gender and Age Distribution | Figure . Future Plans by Course |
| Figure . Workload Satisfaction Analysis    Figure . Percentage Attendance by Course | Figure . Student Time Usage by Course |

**Figure 1** There is a greater spread in male ages and a 3:1 male: female gender ratio.

**Figure 2** Overall 80% will seek paid employment after graduating, 13% plan to go on to further education and the remainder are intent on setting up their own business. Nobody plans to travel. All respondents from the H. Dip. in Data Science intend to seek paid employment. All prospective entrepreneurs are in the Computer Science courses. Masters in Data Analytics has the highest proportion planning to go on to further education.

**Figure 3** As we know, Software Eng./Dev. spend most time in lectures/labs. We see here that they are most keen for a reduction in their workload and curriculum and most eager to do more job preparation. Data Science courses (both Masters and H. Dip.) have the highest overall satisfaction rate (both 78%) with their respective workload.

**Figure 4** 67% of the respondents claim to have attended 90%+ of their lectures. Applied Computer Science has the largest spread and lowest mean of attendance values. Software Eng./Dev. have the highest mean attendance.

**Figure 5** On average, Applied Computer Science spend the most time in lectures/labs and studying. Data Science spend the most time on essays/assignments. Most Applied Computer Science students spend little to no time at jobs/internships.

# Data Imputation

Because of the relatively small number of observations we had gathered, we were loath to drop the observations where data quality issues had been detected (see points 2, 3 and 4 in the previous section) during the Data Analysis phase of the project.

SAS’s **UPCASE** function was used to handle two **Gender** values **in lower case.**Three observations had **invalid values** in the **ranking variables.**A**more involved approach** was needed to deal with this (see the relevant **commentary in the SAS code** for the details of this approach – by searching for all occurrences of the word ‘**imputation**’).

# Data Modelling (Machine Learning)

## Random Forest

The aim of this survey was to analyse important parameters that influence the workload satisfaction of students in their curriculum for different categories and to understand their relationships. After analysing multiple methods for multinomial response, we found that the **Random Forest** **s**upervised machine learning method typically provided good results in terms of modelling, prediction and a good classification rate. Refer to our **SAS code in Appendix B** for how we did that.

**Random Forest** works by constructing a multitude of **decision trees** at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees to get a **more accurate and stable prediction**. The method uses **Loss Reduction Variable Importance** to rank the predictors.

The following table summarises **Random Forest**’s results for our survey’s workload satisfaction-related variables (see **Figure 3**):

|  |  |  |
| --- | --- | --- |
| **Type** | **Top Influencing Factors** | **Least Influencing Factors** |
| Current Curriculum | Age -> Level -> Course -> Gender -> Job | Assign -> Lecture -> Self Study -> Attendance -> Travel |
| Jobs or Internship | Level -> Age -> Course -> Gender -> Assign | Job -> Attendance -> Self Study -> Lecture -> Travel |
| Prep for Future Jobs\_Study | Level -> Age -> Course -> Gender -> Assign | Job -> Self Study -> Lecture -> Attendance -> Travel |
| Overall | Course -> Level -> Gender -> Age -> Job | Assign -> Self Study -> Attendance -> Travel -> Lecture |

As the **Modules** variable was found to be **highly correlated with the Course predictor**, it was determined to be unimportant when **Course** was already in the model. We were surprised to find that other potential predictors, including time spent on **Lectures/Labs (attendance), Essays/Assignments, Self-Study, Part-Time Work or Travel** were **insignificant**, as we had assumed that they would have significantly influenced an individual student’s attitude to his/her course’s **workload**.

|  |  |  |
| --- | --- | --- |
| **Type** | **Misclassification** | **Avg. Sq. Error** |
| Current Curriculum | 0.42 | 0.19 |
| Jobs or Internship | 0.46 | 0.17 |
| Prep for Future Jobs\_Study | 0.48 | 0.25 |
| **Overall Workload** | **0.35** | **0.17** |

The above table shows the overall **misclassification score** and **Mean Square Error** calculated under the “Baseline Fit Statistics” of the **fitted Random Forest model**. In the case of the data in this study, we can see that the model achieved a 35% misclassification rate for overall workload or, in other words, an **accurate classification rate of 65%**. This means that the majority of the sample has been classified correctly.

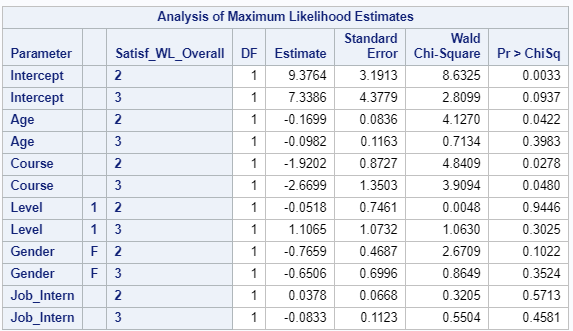
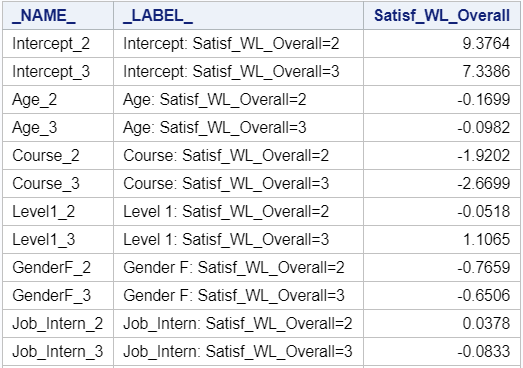
## Logistic Regression

As **Random Forest** is a non-parametric method and, therefore, does not provide a statistical measure of significance for its selected predictors. So, we decided to use **Multinomial Logistic Regression** to interpret the recommended predictors and how they relate to the response variable. It models nominal outcome variables, in which the log odds of the outcomes are modelled as a linear combination of the predictor variables. We tested two models in this multinomial regression, one comparing **Satisfaction Level 1 to 2** and another comparing **Satisfaction Level 1 to 3**, as expressed by the following two equations:

Logit ([**Adequate**]/ [ **I Want Less**]) = β0+β1(Age) + β2(Course) + β3(Level) + β4(Gender) + β5(Job\_Intern)

Logit ([**I Want More**]/ [ **I Want Less**]) = β0`+β1`(Age) + β2`(Course) + β3`(Level) + β4`(Gender) + β5`(Job\_Intern)

Where β’s are the regression estimates shown in table below.

### Satisfaction Level 2-1 (Adequate versus I Want Less)

With P-values < 0.1, **Age, Course and Gender** are rated as being **significant**, and they can be interpreted as follows:

* For one-year increase in Age, we expect to see a 16.9% (OR = 0.844)decrease in the odds their workload being perceived in the category of **Adequate** as compared to **I Want Less**. This means that there are more **Adequately Satisfied** (rather than overworked) younger students compared to their older peers.
* For female students, the odds of workload being in the category of **Adequate** compared to **I Want Less** decreases by 76.6% (OR = 0.216). This means more men are **Adequately Satisfied** (rather than overworked) with their course’s workload compared towomen.
* The relative log odds of being in category **Adequately Satisfied vs I Want Less** will decrease by 1.92 if moving from the level of Course (1-3). This means that most students from the Software Engineering course want less overall workload compared to the other two courses.

### Satisfaction Level 3-1 (I Want More versus I Want Less)

* With P-values < 0.1, only the Course predictor is significant. The relative log odds of being in category I Want More vs I Want Less workload will decrease by 2.67 if moving from the level of Course (1-3). This means that more students from Applied Computer Science want more overall Workload compared to other two courses.

# References

The team found the material discovered in the following websites to be of value during the course of this project:

1. **Survey:** <http://www.ucd.ie/issda/data/eurostudent/>
2. **Graphics:** <https://blogs.sas.com/content/graphicallyspeaking/2017/07/24/lollipop-charts/#prettyPhoto>.
3. **Modelling (Logistic Regression):** <https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/>
4. **Modelling (Random Forest):** <https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d>