**Satisfaction analysis and prediction of Taught Masters Students’ in Ireland**

**Pratik Budhdeo**

2019

# Abstract

*Satisfaction analysis and prediction for Taught Masters Students’ in Ireland*

Educational institutions have a decisive role to play in an internationally growing economy by shaping young citizens to become tomorrow’s workforce and leaders. In this endeavor, the aspect of student’s perceived satisfaction is a critical aspect of their success or failure. Despite that, student voices generally remain unheard of by higher authorities and decision-makers. Using the Eurostudent Survey V data, this thesis presents the findings for student satisfaction prediction, to help in contributing and addressing this information-deficit by generating valuable insights for taught master’s student’s in Irish higher education. This study is based on implementing various machine learning techniques and comparing them to find the best set of predictors. These predictors would then be validated to find the true relationship with the satisfaction, i.e. if it has a positive or negative association and how much they impact satisfaction perception.

On validation, the results revealed that Teaching Quality, Organization of timetable and studies, the Teaching staff's attitude towards students, and Study amenities were the most critical reasons that students consider while rating their satisfaction with college. Satisfaction with accommodation is determined with gender, average monthly expenditure, financial difficulty indicator, felt cheerful indicator and rent paid indicator. Financial difficulty indicator and felt cheerful indicator majorly influence financial and well-being satisfaction.

# Table of Contents

[Abstract 2](#_Toc19987254)

[Table of Contents 3](#_Toc19987255)

[List of Figures 6](#_Toc19987256)

[List of Tables 8](#_Toc19987257)

[Acknowledgments 9](#_Toc19987258)

[Chapter 1 10](#_Toc19987259)

[1.1 Preface 10](#_Toc19987260)

[1.2 Motivation 11](#_Toc19987261)

[1.3 Problem Definition 11](#_Toc19987262)

[1.4 Thesis Structure 12](#_Toc19987263)

[Chapter 2 13](#_Toc19987264)

[2.1 Background 13](#_Toc19987265)

[2.2 Participating Universities 14](#_Toc19987266)

[2.3 Literature Review 15](#_Toc19987267)

[Chapter 3 18](#_Toc19987268)

[3.1 Design 18](#_Toc19987269)

[3.2 Tools and Libraries 18](#_Toc19987270)

[3.2.1 RStudio 18](#_Toc19987271)

[3.2.2 R Libraries 18](#_Toc19987272)

[3.3 Machine learning 19](#_Toc19987273)

[3.3.1 Classical Linear Regression 19](#_Toc19987274)

[3.3.2 Support Vector Machine 19](#_Toc19987275)

[3.3.3 Random Forest 19](#_Toc19987276)

[3.3.4 Linear Discriminant Analysis 20](#_Toc19987277)

[3.3.5 Quadratic Discriminant Analysis 21](#_Toc19987278)

[3.3.6 Multinomial Logistic Regression 21](#_Toc19987279)

[3.3.7 Ordinal Logistic Regression 21](#_Toc19987280)

[Chapter 4 22](#_Toc19987281)

[4.1 Data Management 22](#_Toc19987282)

[4.1.1 Reading Dataset 22](#_Toc19987283)

[4.1.2 Data Cleaning and Imputation 23](#_Toc19987284)

[4.1.3 Data Splitting 24](#_Toc19987285)

[4.2 Exploratory Analysis 25](#_Toc19987286)

[4.3 Satisfaction with College Modeling 27](#_Toc19987287)

[Classical Linear Regression 27](#_Toc19987288)

[Support Vector Machine 29](#_Toc19987289)

[Random Forest 29](#_Toc19987290)

[Linear Discriminant Analysis 31](#_Toc19987291)

[Quadratic Discriminant Analysis 33](#_Toc19987292)

[4.4 Satisfaction with Accommodation Modeling 36](#_Toc19987293)

[Classical Linear Regression 36](#_Toc19987294)

[Support Vector Machine 37](#_Toc19987295)

[Random Forest 38](#_Toc19987296)

[Linear Discriminant Analysis 39](#_Toc19987297)

[Quadratic Discriminant Analysis 42](#_Toc19987298)

[4.5 Satisfaction with Financial/Material wellbeing Modeling 44](#_Toc19987299)

[Classical Linear Regression 44](#_Toc19987300)

[Support Vector Machine 45](#_Toc19987301)

[Random Forest 46](#_Toc19987302)

[Linear Discriminant Analysis 47](#_Toc19987303)

[Quadratic Discriminant Analysis 50](#_Toc19987304)

[Chapter 5 52](#_Toc19987305)

[5.1 Analysis of the predictor-response relationship 52](#_Toc19987306)

[5.1.1 Metrics 52](#_Toc19987307)

[5.2 Satisfaction with the college 55](#_Toc19987308)

[Multinomial Logistic Regression 56](#_Toc19987309)

[Ordinal Logistic Regression 57](#_Toc19987310)

[5.3 Satisfaction with the Accommodation 59](#_Toc19987311)

[Multinomial Logistic Regression 60](#_Toc19987312)

[Ordinal Logistic Regression 62](#_Toc19987313)

[5.4 Satisfaction with the Financial/Material wellbeing 64](#_Toc19987314)

[Multinomial Logistic Regression 65](#_Toc19987315)

[Ordinal Logistic Regression 66](#_Toc19987316)

[Chapter 6 68](#_Toc19987317)

[Conclusion 68](#_Toc19987318)

[Future Work 69](#_Toc19987319)

[References 70](#_Toc19987320)

[Appendix 73](#_Toc19987321)

# List of Figures

[Figure 1 Student satisfaction using survey data and Machine Learning 10](#_Toc19987322)

[Figure 2 Coding Snippet (Imputation) 24](#_Toc19987323)

[Figure 3 Data Splitting 25](#_Toc19987324)

[Figure 4 Satisfaction with College, Accommodation, AND Well-Being 26](#_Toc19987325)

[Figure 5 Exploratory Analysis 26](#_Toc19987326)

[Figure 6 Satisfaction comparison for insignificant predictors (1: - Unsatisfied; 2: - Neutral; 3: - Satisfied) 28](#_Toc19987327)

[Figure 7 SVM (College) 29](#_Toc19987328)

[Figure 8 Random Forest (College) 29](#_Toc19987329)

[Figure 9 Importance Ranking (College) 30](#_Toc19987330)

[Figure 10 LDA (College) 31](#_Toc19987331)

[Figure 11 Partition plot LDA (College) 33](#_Toc19987332)

[Figure 12 QDA (College) 34](#_Toc19987333)

[Figure 13 Partition plot QDA(College) 34](#_Toc19987334)

[Figure 14 SVM (Accommodation) 37](#_Toc19987335)

[Figure 15 Random Forest (Accommodation) 38](#_Toc19987336)

[Figure 16 Importance Ranking (Accommodation) 39](#_Toc19987337)

[Figure 17 LDA (Accommodation) 40](#_Toc19987338)

[Figure 18 Partition plot LDA (Accommodation) 41](#_Toc19987339)

[Figure 19 QDA (Accommodation) 42](#_Toc19987340)

[Figure 20 Partition plot QDA(Accommodation) 43](#_Toc19987341)

[Figure 21 SVM (Wellbeing) 46](#_Toc19987342)

[Figure 22 Random Forest (Wellbeing) 46](#_Toc19987343)

[Figure 23 Importance Ranking (Wellbeing) 47](#_Toc19987344)

[Figure 24 LDA (Wellbeing) 48](#_Toc19987345)

[Figure 25 Partition plot (Wellbeing) 49](#_Toc19987346)

[Figure 26 QDA (Wellbeing) 50](#_Toc19987347)

[Figure 27 Partition plot (Wellbeing) 51](#_Toc19987348)

[Figure 28 Satisfaction with College 55](#_Toc19987349)

[Figure 29 Multinomial Logistic Regression (College) 56](#_Toc19987350)

[Figure 30 Ordinal Logistic Regression (College) 58](#_Toc19987351)

[Figure 31 Satisfaction with Accommodation 59](#_Toc19987352)

[Figure 32 Multinomial Logistic Regression (Accommodation) 60](#_Toc19987353)

[Figure 33 Ordinal Logistic Regression (Accommodation) 62](#_Toc19987354)

[Figure 34 Satisfaction with Financial/Material Wellbeing 64](#_Toc19987355)

[Figure 35 Multinomial Logistic Regression (Wellbeing) 65](#_Toc19987356)

[Figure 36 Ordinal Logistic Regression (Wellbeing) 67](#_Toc19987357)

# List of Tables

[Table 1 Participating Universities 14](#_Toc16255347)

[Table 2 Sample Dataset 22](#_Toc16255348)

[Table 3 Linear Regression coefficients (College) 27](#_Toc16255349)

[Table 4 Significance Evaluation for Predictors (College) 27](#_Toc16255350)

[Table 5 Linear Regression coefficients (Accommodation) 36](#_Toc16255351)

[Table 6 Parameter Significance Evaluation (Accommodation) 36](#_Toc16255352)

[Table 7 Linear Regression coefficients (Wellbeing) 44](#_Toc16255353)

[Table 8 Significance Evaluation (Wellbeing) 44](#_Toc16255354)

[Table 9 Comparison of Models (College) 55](#_Toc16255355)

[Table 10 Comparison of Models (Accommodation) 59](#_Toc16255356)

[Table 11 Comparison of Models (Wellbeing) 64](#_Toc16255357)

# Acknowledgments

I would first like to thank and convey my humble gratitude to my thesis advisor **Dr. Katarina Domijan**, Department of Mathematics & Statistics, NUIM for providing me direction throughout my thesis work. She was very considerate and allowed me to work on a project that was of my personal interest and was always there whenever I had any questions about coding or writing. She gave me continuous feedback and kept me motivated.

Finally, I am sincerely thankful to everyone in the Statistics teaching Department, NUIM. This accomplishment would not have been possible without them.

# Chapter 1

## 1.1 Preface

Irish Higher Education Institutions and universities always strive for excellence in all departments of Teaching & Learning, Research /Innovation, and Knowledge Transfer. Ireland is competing to retain its own and additionally attract international students within a highly competitive international global environment. In the context, not much is known about how they are feeling during their course and its determining factors. [8] This thesis tries to address the research gap, analyzing students’ experiences and predict the reason behind their different levels of satisfaction.



Figure 1 Student satisfaction using survey data and Machine Learning

Figure 1 shows the diagrammatic representation of the research work. Students survey data is analyzed using Machine Learning techniques to predict if they are satisfied(happy) or dissatisfied(unhappy). Using the data obtained from the Eurostudent Survey (2013), I have filtered a subset of data for this thesis. Taught Master’s students pursuing a 1-year or 2-years program in Ireland are included in the analysis. This excludes Postgraduate Cert/Diploma and Research Master’s Students. Using this definition of Taught Master’s Students in Ireland (hereafter ***TMSI***), a sample of 607 students or 6% of the total was retrieved. The total number of participants who completed the survey were 10,110.

The background variables include age, gender country of birth, finances, social standing, parent’s education, etc. Institutional factors included education institution attended, area of study, the reason for choosing a college, etc. Parameters under the wellbeing category include their health assessment, expenses, and peer interaction and relationship. The response variables include students’ perceived satisfaction with their college, accommodation, and well-being.

## 1.2 Motivation

I am a master’s student in Ireland and I often wonder how the perceived satisfaction varies from one individual to another. I had the opportunity to analyze certain aspects of one of my mini-projects that I completed during my second semester. I wondered if I can take it to next level, hence I started looking for ideas and approached my guide to discuss if I can use Machine Learning algorithms to derive important parameters from a given large set of predictors to analyze the student responses.

In this dissertation, I examine survey data and evaluate the aspects of a student’s life that help to relate to their overall contentment. These results might bring us one step closer to solving the problem faced by institutes and governments that are worried about the decreasing rate of overall satisfaction among students with their study duration in Ireland. The challenge to overcome is to select the best ML technique by comparing different metrics.

## 1.3 Problem Definition

Using the survey data, we can study a student’s experience within a learning environment. To provide a solution to the above question, we can use machine learning algorithms. Predicting satisfaction involves validating the governing characteristics and answer questions like: -

* How can we filter the data to fulfill our needs in the best possible way?
* To what extent different characteristics are associated with satisfaction.
* To what extent can we predict satisfaction accurately.

The training model should be able to predict on any given test data to classify the levels of satisfaction. The scale of satisfaction in the original survey data is 1-5 (Very Dissatisfied, Dissatisfied, Neither, Satisfied, Very Satisfied). To improve accuracy and given the size of data, I have re-labeled them to have 3 distinct classes **Dissatisfied**, **Neither** and **Satisfied** by merging the extreme cases.

## 1.4 Thesis Structure

The report explains all tasks carried out from laying the foundation of the project till the completion with six different chapters. These include Introduction, Background, Design and Methodology, Data Management, Results, and Conclusion.

* **Introduction (Chapter 1):** This chapter formalizes the knowledge about student satisfaction and the motivation behind selecting the topic for this project. The main questions that need to be analyzed.
* **Background Work (Chapter 2):** It includes literature review and background research based on the research from previous attempts to predict satisfaction for people from different backgrounds and domains.
* **Design and Methodology (Chapter 3):** Various tools, technologies, and libraries are discussed in this chapter. Also, ML techniques are explained in detail. R Language is primarily used for coding.
* **Implementation (Chapter 4):** It is the heart of the thesis and includes an explanation of data management including reading, understanding, and filtering of data. It highlights the implementation of algorithms and their interpretation complemented by visualization and prediction.
* **Results (Chapter 5):** Various matrices are used for comparison of algorithms to select the best one to obtain the most accurate results for predicting the significant predictors and their relationship to the responses.
* **Conclusion (Chapter 6):** It helps to conclude the analysis and present a detailed summary of the outcome of the project and future scope of enhancements.

# Chapter 2

## 2.1 Background

Ireland was one of 25 nations that took part in the EUROSTUDENT V survey (conducted online in 2013). The student population comprises of undergraduate (4/5th) and postgraduate (1/5th). Higher education is indispensable for Ireland’s to make headway in social and economic progress as there is a clear correlation between the level of education attained and opportunities for employment. There are noteworthy development and expansion of student’s participation in Ireland over recent decades (15).

Student satisfaction is assumed to be a vital feature of academics as a monitoring indicator of excellence and quality. Primarily, linking quality and satisfaction is a method often used to improve outcomes in different fields (2). Carney (4), for example, proposes various predictors that influence college and institutes, like academic reputation, the structure of curriculum, mean class capacity, their scholastic qualities, ease of communication between the staff and the students, students’ social life, tuition fees, etc. While conversely, Athiyaman (2) articulates the relevance of the library, teaching quality, health and wellbeing of students, counseling options, technical facilities, and also course content difficulty levels and student workload. The background research suggests that there is still scope of a distinguishing the concrete association between the predictors to the perceived student’s satisfaction in the colleges.

## 2.2 Participating Universities

There is a total of twenty-six Irish higher education institutes as participants of this survey. See Table 1.

|  |  |  |
| --- | --- | --- |
| **Type** | **Institute** | **Students** |
| Institutes of Technology (IT) | Dublin IT | 40 |
| Cork IT | 18 |
| Waterford IT | 18 |
| Carlow IT | 10 |
| Athlone IT | 9 |
| IT, Tralee | 5 |
| IT, Tallaght | 3 |
| Dunlaoghaire Institute of Art, Design, and Technology | 2 |
| Letterkenny IT | 2 |
| Limerick IT | 2 |
| Galway-Mayo IT | 1 |
| IT, Blanchardstown | 1 |
| IT, Sligo | 1 |
| Dundalk IT | 0 |
| Other | Mater Dei Institute of Education | 8 |
| National College of Art and Design | 5 |
| St. Patrick's College Drumcondra | 4 |
| Mary Immaculate College | 3 |
| St. Angela's College of Education | 1 |
| Universities | University College Dublin (UCD) | 170 |
| National University of Ireland, Galway (NUIG) | 93 |
| Trinity College Dublin (TCD) | 74 |
| National University of Ireland, Maynooth (NUIM) | 47 |
| University College Cork | 46 |
| Dublin City University | 45 |
| University of Limerick | 41 |

Table 1 Participating Universities

## 2.3 Literature Review

As outlined in previous chapters, I have explained the motivation and background behind selecting the topic for the research work. In continuation, I will cover the associated work that has been done in the past and was used as the foundation stone for this project. Though there have been previous researches’ in this domain, they have been from a standalone perspective where only 1 algorithm was chosen at random and they were not cross-validated against other possible modeling techniques that could have performed better. The reason could be that these researches were not conducted by statisticians and instead managed by experts in other domains without a thorough knowledge of possibilities provided by machine learning algorithms.

**Duque, L. C. (2014). [7]**

Duque tried to determine a conceptual system to understand three things. Firstly, students' satisfaction with college. Secondly, perceived learning outcomes on course completion. Most importantly, intentions of dropout of college in case their goals are not met or otherwise. The aim of the study was to better understand how these perceptions influence their future intentions. These data sets were from three countries across various undergraduate programs. The questionnaire was designed to be adaptive to specific contexts. The different algorithms were verified using equation modeling. A significant finding was that participants' co-creation (student-led collaborating initiatives) was as significant as perceived service quality in justifying their intellectual results from learning, thus explaining higher-level effective learning outcomes.

**Eurostudent V: Ireland 2013 [8]**

Eurostudent Survey V report discusses the descriptive analyses of the 5th Eurostudent survey of HEI in Ireland. This report tries to analyze the learning experience of the students and understand their life-quality and its influencing factors. Secondly, it tries to list the challenges faced by institutions in ensuring a high-quality learning environment for all the students. The report uses basic statistics and visualizations but does not include any machine learning implementations to analyze the significance of parameters on perceived satisfaction.

**European Commission/EACEA/Eurydice, 2018. [9]**

This report was set up for the EMC in Paris, in May 2018. It reflected various perspectives of the state of implementation of the Bologna Process. Eurostudent VI survey data collected in 2017 was used to validate the economic and social conditions of participants. Along with other objectives, it explored the impact of teaching quality from a student perspective on their individual satisfaction. The survey shows that satisfaction is overall quite high. Ireland ranked 5th among all countries in terms of the highest level of satisfaction – 71 % of students satisfied or very satisfied with teaching quality. Only descriptive analysis was included without the implementation of any machine learning techniques.

**El Ansari W. (2004) [10]**

Ansariinspected 3 categories with an aim to increase research knowledge of health professionals. The impact of geographic and institutional predictors was investigated in relation to the student contentment. The reliability of latent scale variables was analyzed by Cronbach’s Alpha reliability coefficient where outcomes more than 7/10th were viewed as reliable. As demonstrated by the outcome, there was notable contrast in satisfaction with the various aspects of modules like content, integration, administration, assessment, and stimulation.

**Finn, M. and Darmody, M. (2016). [11]**

This paper examines the students studying in Irish HEI not belonging to Ireland for their overall experiences. The study was based on weighted data to eliminate bias in the outcome. It assesses the factors to incorporate in the multivariate modeling technique. The logistic regression technique helped to predict the relationship among different variables. The result suggested that the most critical aspect influencing student satisfaction with the study was the institution in which they were studying.

**Jokar, Saeid. (2018). [18]**

This study analyzed the student satisfaction from the skills workshop (research domain) in Iran conducted in 2011 at Yasuj University by the southern network of the student research committee. Information was gathered using a two-part survey. The initial part comprised of demographic questions and the later included 20 closed questions related to participant satisfaction. Information was investigated with SPSS utilizing an independent T-test. In this investigation web (internet) inaccessibility was the primarily significant constraint of this workshop, subsequently, a very high percentage (70.5%) of the students dis-satisfied with the conduct of the workshop.

**Onditi, E. O., &Wechuli, T.W. (2017). [23]**

This paper examined administration quality and satisfaction for HEI students from different sub-categories and streams. It analyses the Gap method and the hierarchical model of service quality assessment. The outcomes indicated that there was no agreement among different authors (who had previously conducted similar researches’) on the capacity that should be checked to verify the service quality. This paper thus emphasized that institutions ought to create mechanisms to gather student feedback with the goal that they can make the necessary enhancements for the critical administration quality measurements.

**Skrbinjek, V. & Dermol, V. Tert Educ Manag (2019) [27]**

This research underlined students’ satisfaction and its relationship to their presentation and inclusion in different exercises in the e-study classroom in Slovenia. Multiple linear regression analysis followed by decision-tree (predictor variables were the average students’ marks, number of students, average workload and involvement of students in the e-classroom) was used for modeling. The outcomes uncovered that understudies are more unhappy with the curriculum when both the attendance requirement in the e-study and the cumulative pressure are higher. It also flagged that the mean grade of the class probably won't be pivotal when analyzing a student’s perceived satisfaction.

# Chapter 3

## 3.1 Design

This chapter covers about design, tools and methodologies used in this project. The first part reason for selecting R as the language for all coding involved and the libraries that helped to implement the machine learning algorithms. In the latter part, I have included a basic theoretical description of all the algorithms starting from classical linear regression to complex techniques like Random Forest and LDA.

## 3.2 Tools and Libraries

### 3.2.1 RStudio

I have used the RStudio platform for all coding and analysis involved in this thesis. This software is an open-source and free integrated development environment (IDE) for R, used for statistical analysis and visualization.

Though Python is a general-purpose language with an easy-to-read syntax, R's core functionality is developed while keeping statisticians on the front, thereby it has specific great features for data visualization and modeling. [27] R Language provides an immense number of useful libraries for data manipulation and outcome prediction.

### 3.2.2 R Libraries

R provides a pool of different inbuilt libraries that help to make the code simpler and implement algorithms, some important ones are “**stats”** (28) (Basic manipulation and lm function), “**haven**” (14) (read\_sav function for reading sav files from SAS), “**nnet**” (30) (multinom model), “**MASS**” (30) (Polr, lda and qda implementation), “**e1071”** (5) (svm functions), “**randomForest**” (1) (randomForest implementation),”**skimr**”(21) (skim\_to\_wide for descriptive statistics),” **tidyverse**” (13) (ggplot2),”**klaR**”(31) (lda plots).

## 3.3 Machine learning

Shalev and Ben [26] have suggested that the science of discovering the useful patterns in the data set is known as machine learning. It is the art of making computers act and react without any kind of explicit programming. I have summarised all the techniques that are used for analysis and result interpretation.

### 3.3.1 Classical Linear Regression

It is used to model the association between a dependent variable and one or more explanatory variables. It consists of the mean function and the variance function. [32]. In many techniques, predictors are not known and must be predicted using data. The data is assumed to have a normal distribution and prediction errors can be minimized to increase accuracy.

“lm” function is used from the stats (28) package.

### 3.3.2 Support Vector Machine

SVM is an extended version of the SVC (classifier). It allows for non-linear classification boundaries (hyperplane or decision) using a general technique known as kernels where different kernels are introduced to decide the linearity. Separating the hyperplane is not possible as there might be values placed in the incorrect classes. The model permits this violation to some extent if it is useful for minimizing the overall misclassification. This is regulated by tuning parameter C (6). “svm” function inside “e1071” (5) package is used.

### 3.3.3 Random Forest

This algorithm enables us to extract the most significant predictors in a vast dataset over the input variables by producing numerous decision trees and then ranking the variables by importance (20). During training, each tree gains knowledge from a random sample of observations drawn with substitution, known as bootstrapping (20). It significantly improves the bias-variance tradeoff. The Gini index G is often used instead of classification error. The value lies between 0 (perfect classifications) and increases as the node becomes more impure. The importance of each predictor can be estimated by the mean value of the Gini index decrease from the partitions from the forest. A bigger value suggests that the predictor is relevant (significant). Using this information, the Variable Importance plot can be generated by placing predictors in decreasing order of importance. “randomForest” function inside “randomForest” package (1) is used for analysis.

Model tuning is a method that helps to find the best settings for an ML algorithm by varying below: -

* No. of decision trees
* Maximum depth of each decision tree
* The highest number of predictors used for splitting each node
* The number of data points permissible in a leaf node.

### 3.3.4 Linear Discriminant Analysis

The technique for recognition of patterns to extract a combination (linear) of predictors that categorizes two or more classes of response. The results can help in reducing the dimensionality from a vast set of feature space.

It assumes that there is a normal distribution predictor space and the classes have identical covariance matrices (for multivariate analysis, p > 1) (3). LDA calculates “discriminant scores” of each point to classify what response variable category it is in (i.e. default or not default).

From the summary, we get below information: -

The **Prior probabilities of groups** (πi) - Probability of selecting an observation from class “i” at random amongst the complete dataset.

The**Group means** (μi) - Average for each variable for each class i that is independent.

The **Coefficients of LD** - Coefficients for each linear discriminant.

The **Proportions of the trace** - Ratio of between-class variance calculated by successive discriminant functions. “lda” function inside the “Mass” package (30) is used for analysis.

Partition plots yield a series of plots for all combinations of available categories. Each plot is a different version of the same dataset. Shaded regions depict different predicted classes. **Unsatisfied** represented by **blue,** **Neutral** by **white** and **Satisfied** by **pink** region. The apparent error rate is calculated for each view and cascaded above the plots. The function plots the ability of the predictors to partition the response class taking combinations of two at a time using a similar technique as lda. “partimat” inside klaR (31) package is used for generating plots.

### 3.3.5 Quadratic Discriminant Analysis

QDA is more flexible than LDA as it relaxes the assumption of the equality of variance/covariance of the predictor space. The distributions, i.e. each class is allowed have a different covariance structure (3).

“qda” function inside the “Mass” package (30) is used for analysis.

### 3.3.6 Multinomial Logistic Regression

MLR is a technique used to analyze a discrete outcome response with more than two classes. The log odds are abstracted as a linear combination of the predictor variable space [17].

“nnet” package (30) is used to build a “multinom” model.

### 3.3.7 Ordinal Logistic Regression

OLR is a regression technique used for predicting an ordinal response, i.e. variable whose value exists such that only the successive ordering among the classes is relevant. This method is based on an adaptation of the multinomial likelihood (17).

“Polr” function inside the “Mass” (30) package is used for analysis.

# Chapter 4

## 4.1 Data Management

Data Management is analyzing the data and organizing it to find the solution in the best possible way. I have explained the data retrieval and filtering; data splitting techniques used to obtain the final subset used in this thesis. Exploratory analysis and implementation of the various techniques described in the previous chapter are a part of this chapter. This aim is to find the best technique for classification of different levels of satisfaction.

### 4.1.1 Reading Dataset

The dataset is Eurostudent Survey V is read in R using *read\_sav format.* It is filtered based on Study Program as “Taught Masters” and Course Duration as “1 or 2 years” using *subset* function to create a dataset “Taught Master’s Students in Ireland” TMSI. We get 607 students. The snapshot of the dataset is in Table 2.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | Q1.2 | Q1.4 | Q1.5 | Q3.7 | Q3.8.1c1 | Q3.8.1c2 | Q3.12 | Q3.14 | Q5.1 |
| 1 | 2 | 22 | 3 | 400 | 200 | 0 | 2 | 0 | 55.91663 |
| 2 | 2 | 26 | 1 | 150 | 2090 | 0 | 3 | 40 | 39.49997 |
| 3 | 1 | 23 | 10 | 800 | 565 | 150 | 4 | 20 | 27.5833 |
| 4 | 2 | 18 | 8 | 1300 | 1250 | 0 | 3 | 30 | 26.8333 |
| 5 | 1 | 18 | 3 | 120 | 110 | 310 | 3 | 0 | 24.3333 |
| 6 | 1 | 18 | 6 | 0 | 0 | 647.5 | 5 | 0 | 24.41663 |
| 7 | 2 | 3 | 3 | 752 | 160 | 70 | 4 | 20 | 28.91663 |
| 8 | 1 | 25 | 11 | 0 | 1040 | 0 | 5 | 0 | 37.66663 |
| 9 | 1 | 24 | 2 | 390 | 295 | 60 | 4 | 10 | 29.41663 |
| 10 | 2 | 24 | 3 | 2643 | 1065 | 0 | 3 | 35 | 40.16663 |
| 11 | 1 | 3 | 2 | 0 | 775 | 0 | 5 | 16 | 30.5833 |
| 12 | 1 | 18 | 5 | 417 | 415 | 450 | 2 | 0 | 21.99997 |
| 13 | 1 | 24 | 6 | 0 | 1295 | 0 | 1 | 40 | 31.5833 |
| 14 | 1 | 18 | 2 | 310 | 410 | 0 | 2 | 0 | 23.24997 |
| 15 | 2 | 25 | 4 | 150 | 1120 | 0 | 4 | 26 | 27.49997 |
| 16 | 1 | 4 | 8 | 500 | 0 | 410 | 5 | 0 | 21.91663 |
| 17 | 1 | 18 | 7 | 500 | 320 | 300 | 4 | 10 | 21.74997 |

Table 2 Sample Dataset

***Predictors*** are divided into 3 categories as below: -

* **Background -** Age, gender, country of birth, finances, social standing, parent’s education and profession, and family income.
* **Institutional -** Educational institution attended, area of study, the reason for selecting a college, and accommodation.
* **Well-being** - Perceived assessment of health, lifestyle and financial standing.

***The response*** includes Students’ perceived satisfaction with: -

* College
* Accommodation
* Financial and material Wellbeing

### 4.1.2 Data Cleaning and Imputation

The subsequent step after data gathering is finding the inconsistencies in the data set before the actual implementation of algorithms. This method is defined as data cleaning (24). Missing values are viewed as the first obstacle in data analysis. The decision to select a technique to fix missing has a large impact on the model precision.

As a default approach, most techniques perform listwise deletion to fix unaccounted entries in a row even if only a single value is missing in the entire row that may / may not be important for modeling. Hence it leads to information loss and reduced model accuracy. Skimming helps us to analyze the number of missing values in the complete dataset at once. “skim\_to\_wide” function inside skimr package is used to all list missing values at once.

Missing data in this survey questionnaire was handled with separate caution in each case and appropriate imputation techniques were implemented. This includes: -

* Removed – If missing at random and can’t be treated/analyzed.
* Recovery – The survey demanded specific values.
* Imputation using neighbor’s knowledge – When the proportion of missing values was low.

Code snippet for variables (sample) that were imputed are shown in Figure 2: -

* Student category (Mature or not) – Updated based on age.
* Monthly disposable income – Survey form required values to be marked as 0 instead of NA.
* Irish or Non-Irish – Based on the country of birth and citizenship information.

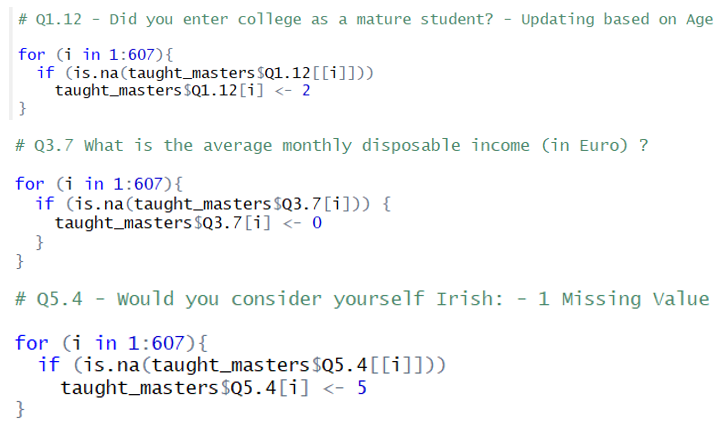


Figure Coding Snippet (Imputation)

### 4.1.3 Data Splitting

From the overall data, I have set aside 33% testing data which would be used later to help predict the accuracy of the model. A model that fits well on data that is not used for training is an efficient one, also it helps to reduce the bias. We divide the data as below: -

* **Training**: Points whose labels are used to classify unlabeled points. (67%)
* **Test**: Points not used as labeled cases, used to estimate the classification error of our fitted model. (33%)

Below are the methods that can be useful in treating imbalanced datasets:

1. Under-sampling
2. Oversampling
3. Synthetic Data Generation
4. Cost-Sensitive Learning

“createDataPartition” function inside the caret package [19] is used for created different splits based on the response classification. In this function, we can select the response variable and split proportion. With the help of metrics comparison, best results were obtained with a partition of 2:1 with training and testing set.

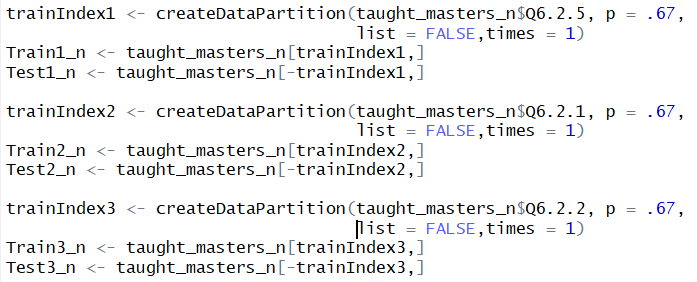


Figure 3 Data Splitting

## 4.2 Exploratory Analysis

In the original survey, the satisfaction scale is 1-5. I have re-grouped them to a scale of 1-3 as some of the groups had limited responses (unbalanced data). The three groups are as below: -

**1: - Dissatisfied; 2: - Neutral; 3: - Satisfied**

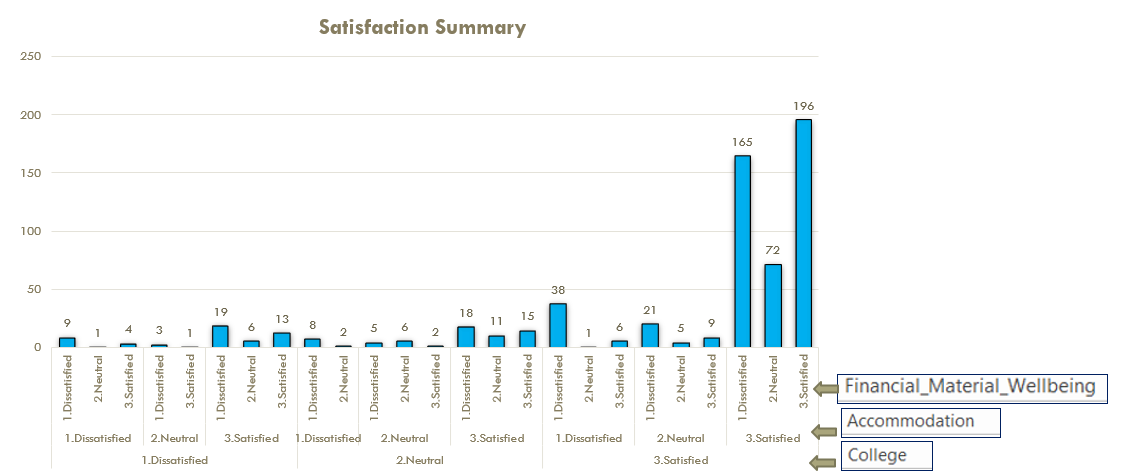


Figure 4 Satisfaction with College, Accommodation, AND Well-Being

In Figure 4, we can see that most students are satisfied with college, but the reverse is true in the case of Well-being. This is a crucial aspect where the most focus needs to be provided by the colleges and government to make the situation better.

Figure 5 Exploratory Analysis

The characteristics of the participants are analyzed across factors including gender, maturity and course duration. [Figure 5] The gender-split of students shows 1.5 times more male participants than female. Approximately 56% of all students are mature. The one-year course is more popular and is chosen by 61% of students which shows that it is unarguably the first choice of students.

## 4.3 Satisfaction with College Modeling

### Classical Linear Regression

Classical regression analysis helps to understand the relationship between dependent and independent factors. But in our case the data isn’t normally distributed and has a non-linear shape, thus findings of the analysis are indecisive and not used in results formulation and model comparison at a later stage.

Coefficients:

Estimate Std. Error t value PR(>|t|)

(Intercept) 0.666086 0.160402 4.153 3.77e-05 \*\*\*

Q1.4 -0.001453 0.002647 -0.549 0.58331

Q1.14.1 0.187572 0.034562 5.427 8.35e-08 \*\*\*

Q1.14.2 0.051689 0.023613 2.189 0.02899 \*

Q1.14.3 0.001061 0.022871 0.046 0.96303

Q1.14.4 0.048869 0.022809 2.143 0.03256 \*

Q1.14.5 0.147623 0.028515 5.177 3.09e-07 \*\*\*

Q1.14.6 0.067907 0.021742 3.123 0.00188 \*\*

Q1.12 0.015669 0.041925 0.374 0.70873

Q5.2 0.033900 0.041688 0.813 0.41645

Q5.42 -0.057316 0.051920 -1.104 0.27007

Q7.52 0.011777 0.048053 0.245 0.80648

Q7.53 0.018706 0.095252 0.196 0.84438

---

Signify. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘’ 1

Table 3 Linear Regression coefficients (College)

|  |  |
| --- | --- |
| **Predictor** | **Significant** |
| Q1.4 - College attending? | **N** |
| Q1.12 – Are you a mature student? | **N** |
| Q5.2 - e\_gender | **N** |
| Q5.4 - Would you consider yourself: Irish or Non-Irish | **N** |
| Q7.5 - Social Class | **N** |
| Q1.14.4 - College administration's attitude | **N** |
| Q1.14.1 - Teaching quality | **Y** |
| Q1.14.2 - Organization of timetable and studies | **Y** |
| Q1.14.5 - Staff's attitude towards students | **Y** |
| Q1.14.6 - Study amenities (e.g. computers, library, classrooms) | **Y** |

Table 4 Significance Evaluation for Predictors (College)

#### Interpretation

The response is “Students perceived satisfaction with college” and significant predictors are derived from Table 3, where P-Value<0.05**,** i.e. in our case Satisfaction with *Organisation of timetable and studies, Teaching Quality, Teaching staff's attitude as well as amenities* (e.g. computers and library). For all the significant predictors in Table 4, **coefficients are positive**, i.e. Satisfaction with college increases with an increase in satisfaction with the predictors.

Some of the predictors which were assumed to be significant before the analysis on a general perception were found to be insignificant. Factors like **age, gender, perceived social class and maturity were marked as not significant**; i.e. they do not impact the overall student satisfaction. This can be visualized using Figure 6.

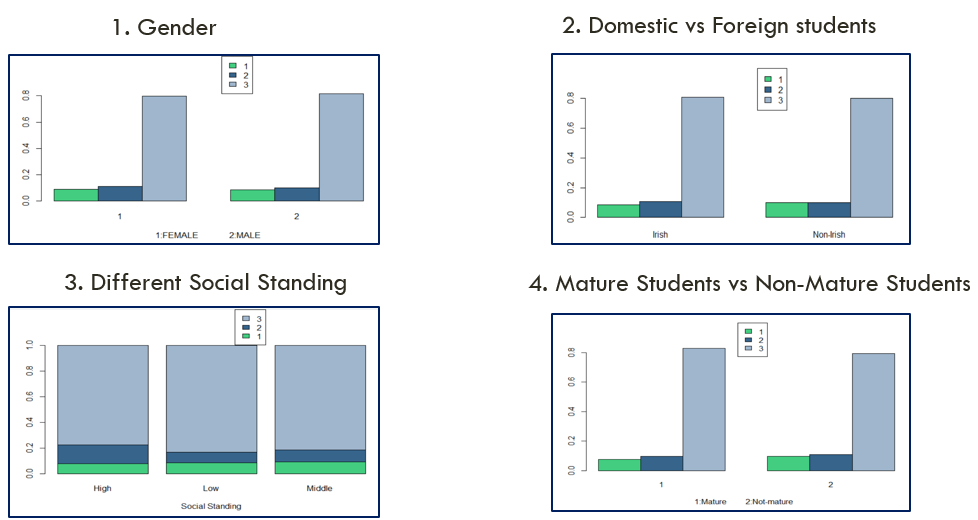


Figure 6 Satisfaction comparison for insignificant predictors (1: - Unsatisfied; 2: - Neutral; 3: - Satisfied)

These bar plots help us to visualize factors for which the satisfaction is the same for all classes of response, hence they do not provide any valuable insight into the models and are not significant in predicting the student’s satisfaction of college.

### Support Vector Machine

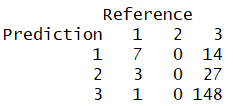
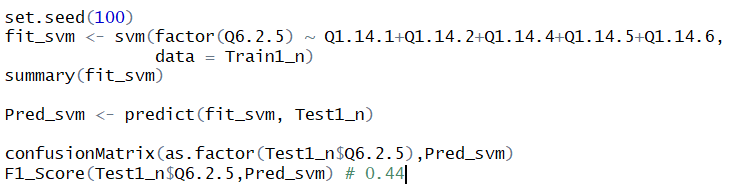
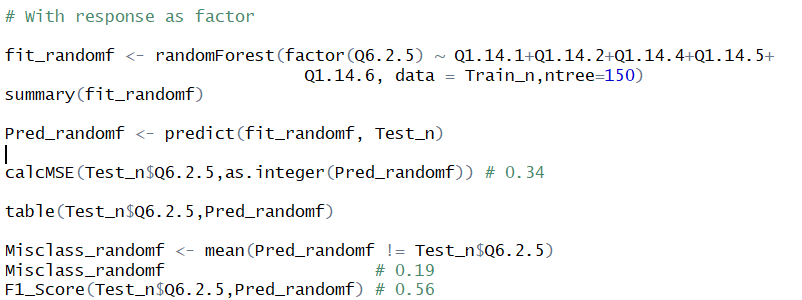


Figure 7 SVM (College)

#### Interpretation

I have used the multi-category support vector machine algorithm fromthee1071 package (5). **SVM** fits the test data correctly and misclassified only 22.5% observations. Again, in this case, Group 2 (“Neutral” in terms of satisfaction with the college) is misclassified 100%. Overall accuracy is 77.5% and it does a great job in predicting “Unsatisfied” and “Satisfied” students.

### Random Forest



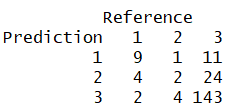


Figure 8 Random Forest (College)

#### Interpretation

**Random Forest** fits the test data correctly and misclassified only 19% observations.

Again, in this case, Group 2 (“Neutral” in terms of satisfaction with the college) is misclassified 100%. Overall accuracy is 81% and it does a great job in predicting “Unsatisfied” and “Satisfied” students.

#### Visualization

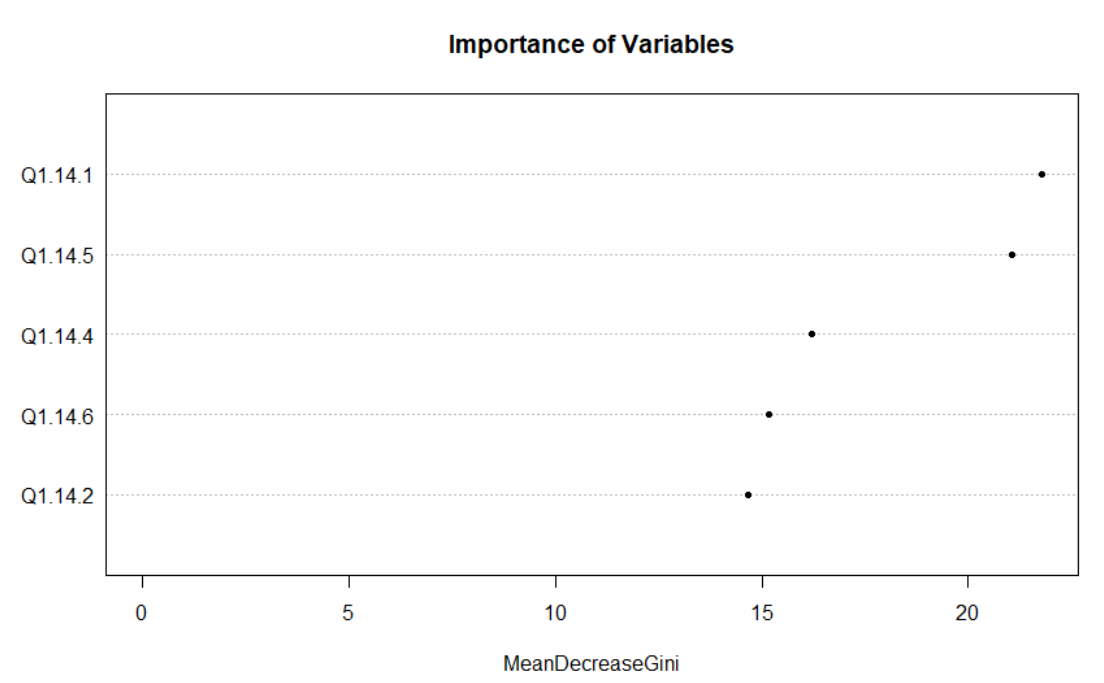


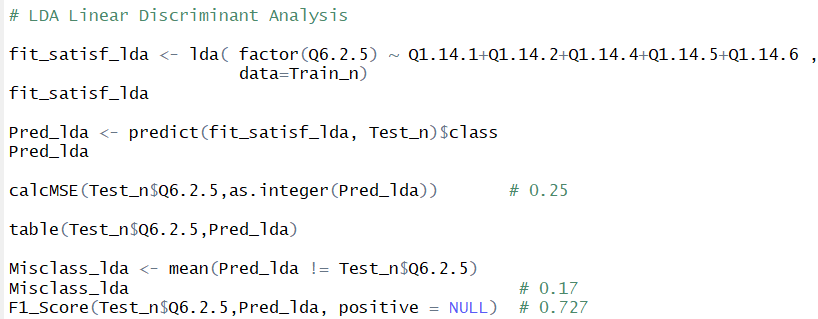
Figure 9 Importance Ranking (College)

Figure 9 depicts “Importance ranking” for predictors: -

1. Quality of teaching (Q1.14.1)
2. Teaching staff's behavior towards students (Q1.14.5)
3. College administration's behavior towards students (Q1.14.4)
4. Study amenities (e.g. library, computers, classrooms) (Q1.14.6)
5. Organization of timetable and studies (Q1.14.2)

### Linear Discriminant Analysis

LDA has an assumption that there is a normal distribution in data among predictor space. As stated earlier that our data isn’t normally distributed, but we are still interested in the prediction accuracy, hence we implement LDA.



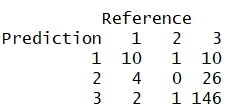
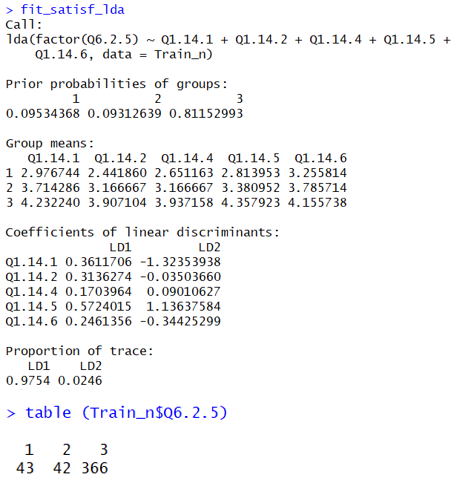


Figure 10 LDA (College)

#### Interpretation

There are 451 observations, 43 in class 1, 42 in class 2 and 366 in class 3. LD1 is the first linear discriminant which is a linear combination of below parameters: -

**(Q1.14.1\*0.3611706) + (Q1.14.2\*0.3136274) + (Q1.14.4\*0.1703964) + (Q1.14.5\*0.5724015) + (Q1.14.6\*0.2461356)**

Teaching staff's behavior towards students, Teaching Quality and Organisation of timetable and studies have the largest coefficients which signify them as the most important parameters in the estimation of response.

All the coefficients are positive, this resonates with the result that increases in satisfaction of any of the included predictors in the model lead to an overall increase in student satisfaction with college.

**LD1 explains 97.54% of the overall variance as** shown in Figure 10.

#### Visualization

With the help Partition plot (Figure 11), we can visualize how the data among different classes are grouped together.

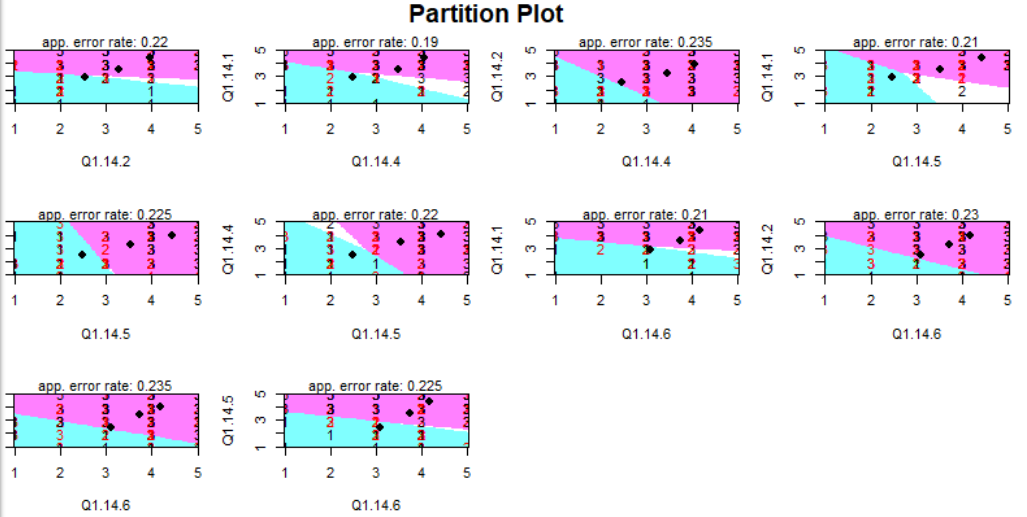


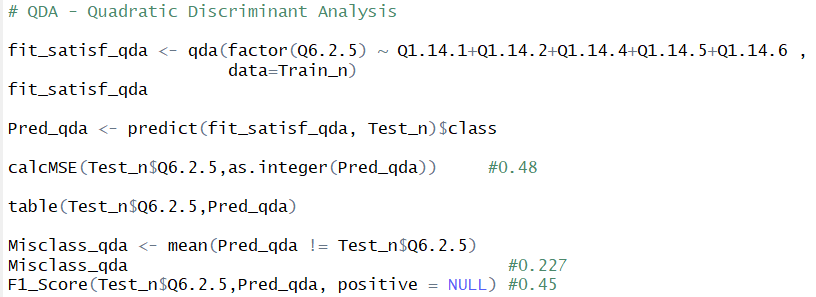
Figure 11 Partition plot LDA (College)

Partition plots are a series of plots for all combinations of categories. The plot shows the three groups with overlapping boundaries. Figure 11 shows that Group 1 and 3 are distinguishable, but Group 2 has major overlapping with the boundary of both the remaining groups and hence it is difficult to correctly predict if a student is “Neutral” in terms of satisfaction with the college.

#### Predictions

This model fit the test data correctly and misclassified only 17% observations. But it doesn’t predict group 2 correctly as there is a lot overlap between the adjacent classes hence it is very difficult to predict if a student is “Neutral” in terms of satisfaction with the college. Otherwise, it does a great job of predicting “**Unsatisfied” or “Satisfied”** students accurately. The model performs comprehensively well with the test data (not used for modeling) having an overall classification accuracy of 83%.

### Quadratic Discriminant Analysis



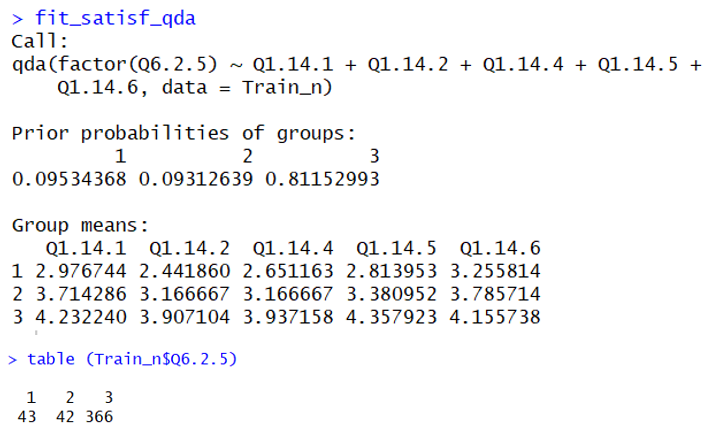
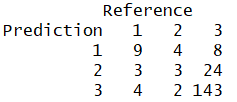
 

Figure 12 QDA (College)

#### Visualization

With the help of the Projection plot (Figure 13) for QDA, we can visualize how the data among different classes are grouped together.

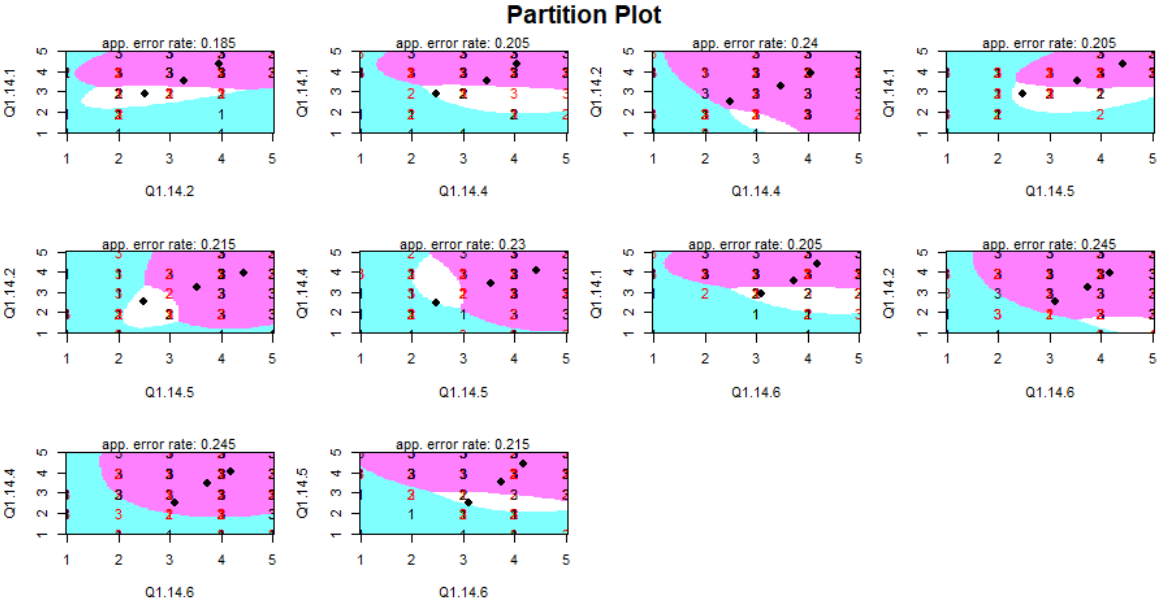


Figure 13 Partition plot QDA(College)

Once again, a series of plots for each combination of three variables is generated where each one is a distinct version of the same data. Shaded regions depict different classes like LDA, but this time boundaries are non-linear. Group 1 and 3 are distinguishable, but Group 2 has an overlapping boundary with remaining groups and hence it is difficult to correctly predict if a student is “Neutral” in terms of satisfaction with the college.

#### Predictions

There are 451 observations, 43 in class 1, 42 in class 2 and 366 in class 3. QDA model predicts the test data correctly but misclassified slightly more that LDA with 22.7% observations. Again Group 2 (“Neutral” in terms of satisfaction with the college) is incorrectly predicted 100% times. Overall it does a fair job in predicting “Unsatisfied” or “Satisfied” students accurately with an accuracy of 77.3%, but not as good as LDA.

## 4.4 Satisfaction with Accommodation Modeling

### Classical Linear Regression

Classical regression analysis helps to understand the relationship between dependent and independent factors. But in our case the data is not normally distributed and has a non-linear shape, thus findings of the analysis are indecisive and not used in results formulation and model comparison at a later stage.

Coefficients:

Estimate Std. Error t value PR(>|t|)

(Intercept) 2.622e+00 2.656e-01 9.870 < 2e-16 \*\*\*

Q3.8.1c1 1.012e-04 5.039e-05 2.008 0.045323 \*

Q6.1.1 7.728e-02 2.590e-02 2.984 0.003021 \*\*

Q3.12 -9.103e-02 2.664e-02 -3.417 0.000698 \*\*\*

Q5.22 -2.032e-01 6.447e-02 -3.152 0.001745 \*\*

Q3.6.2 3.586e-04 9.834e-04 0.365 0.715579

Q3.32 4.662e-02 7.398e-02 0.630 0.528996

Q1.12 -2.048e-02 7.527e-02 -0.272 0.785725

Q3.5.2 2.161e-02 2.322e-02 0.931 0.352459

Q5.1 3.626e-04 4.695e-03 0.077 0.938474

Table Linear Regression coefficients (Accommodation)

|  |  |
| --- | --- |
| **Predictor** | **Significant** |
| Q3.5.2 - What is the most frequent mode of transport? | **N** |
| Q3.3 - Rent for accommodation indicator? | **N** |
| Q3.6.2 - Distance you home to HE | **N** |
| Q5.1 - What age are you (years)? | **N** |
| Q5.2 - e\_gender | **Y** |
| Q3.12 - Experiencing financial difficulties indicator? | **Y** |
| Q6.1.1 - Felt in good spirits and cheerful indicator (over last 2 weeks) | **Y** |
| Q3.8.1c1 - Total - Own Pocket | **Y** |

Table 6 Parameter Significance Evaluation (Accommodation)

#### Interpretation

Table 5 shows how the students perceived satisfaction with accommodation (response) is derived from predictors with P-Value<0.05. Summarised in Table 6 as well.

* *Gender [1: Female and 2: Male]* - Coefficient is negative which signifies that satisfaction level is less for a male when compared to female students.
* *Perceived level of financial difficulties* – Negative coefficient signifies that satisfaction level decreases when students are facing more financial difficulty.
* *Feeling of being in good spirits and cheerful over the last two weeks* - the Positive coefficient signifies that satisfaction level increases when students have felt more cheerful on a scale of 1-6 (At no Time - All the Time).
* *Average monthly expenditure (Own Pocket)* - The positive coefficient signifies that satisfaction level increases when students ' expenditure increases. This is a bit out of the box, and we need to re-validate this outcome when we derive final conclusions after comparing other models.

Again, some of the predictors which were assumed to be significant before the analysis on a general perception were found to be insignificant. Factors like **age, distance traveled, rented accommodation and transport mode were marked as insignificant**; i.e. they do not impact the overall student satisfaction. These predictors do not provide substantial insights into the model and hence marked as not significant in predicting the student’s satisfaction of accommodation.

### Support Vector Machine

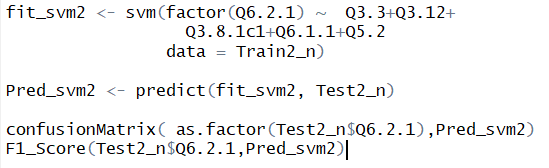
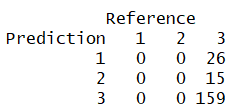
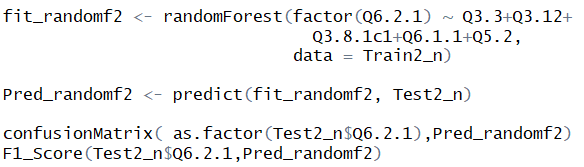
 

Figure 14 SVM (Accommodation)

#### Interpretation

I have used the multi-category support vector machine algorithm fromthee1071 package (5). **SVM** fits the test data correctly and misclassified only 20% observations. In this case, Group 1 and 2 are misclassified 100% of the time. Hence it is not a good model for satisfaction prediction with accommodation for desired reasons.

### Random Forest



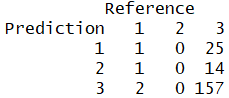


Figure 15 Random Forest (Accommodation)

#### Interpretation

**Random Forest** fits the test data correctly and misclassified only 21% observations. In this case, Group 2 is misclassified 100% of the time. Overall accuracy is 79% and it does a great job in predicting “Unsatisfied” and “Satisfied” student categories for satisfaction with the accommodation.

#### Visualization

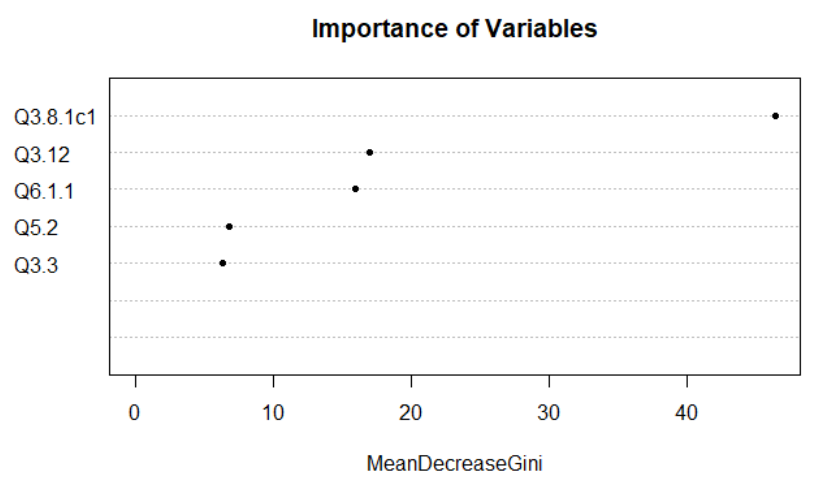


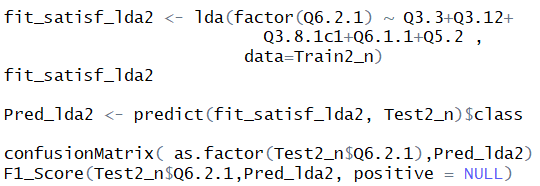
Figure 16 Importance Ranking (Accommodation)

Importance ranking for predictors as below: -

1. Q3.8.1c1 - Total - Own Pocket
2. Q3.12 – Level of Financial difficulties indicator?
3. Q6.1.1 - Felt cheerful indicator (over last two weeks)
4. Q5.2 - e\_gender
5. Q3.3 - Rent for accommodation indicator?

### Linear Discriminant Analysis

LDA has an assumption that there is a normal distribution in data among predictor space. As stated earlier that our data isn’t normally distributed, but we are still interested in the prediction accuracy, hence we implement LDA.



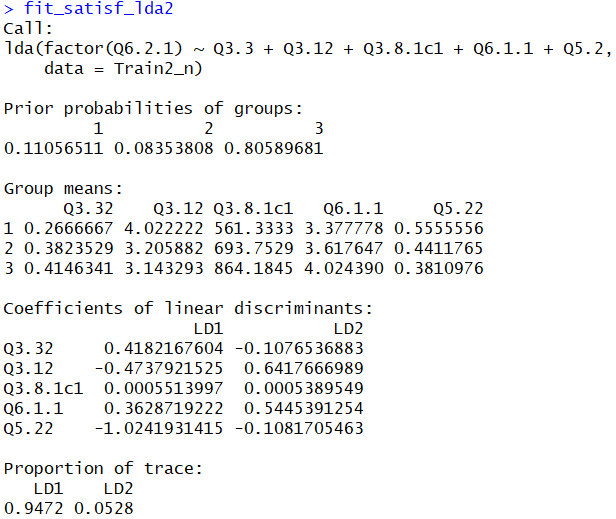
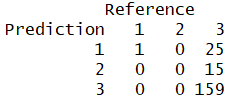
 

Figure 17 LDA (Accommodation)

#### Interpretation

There are 407 total observations, 45 in class 1, 34 in class 2 and 328 in class 3.

LD1 is the first linear discriminant which is a linear combination of below parameters: -

**(Q3.32\* 0.4182167604 + Q3.12\*-0.4737921525 + Q3.8.1c1\* 0.0005513997 + Q6.1.1\* 0.3628719222 + Q5.22\*-1.0241931415)**

Q5.22, Q3.32, and Q3.12 have the largest coefficients, hence these are the most significant predictors in influencing the response.

Q3.12 and Q5.22 have negative coefficients and the rest of the parameters have positive coefficients, this resonates with the result from linear regression for predicting student satisfaction with the accommodation. **LD1 explains 94.72% of the overall variance in the data**.

#### Visualization

Utilizing the Partition plot [Figure 18], we can visualize how the data among different classes are grouped together.

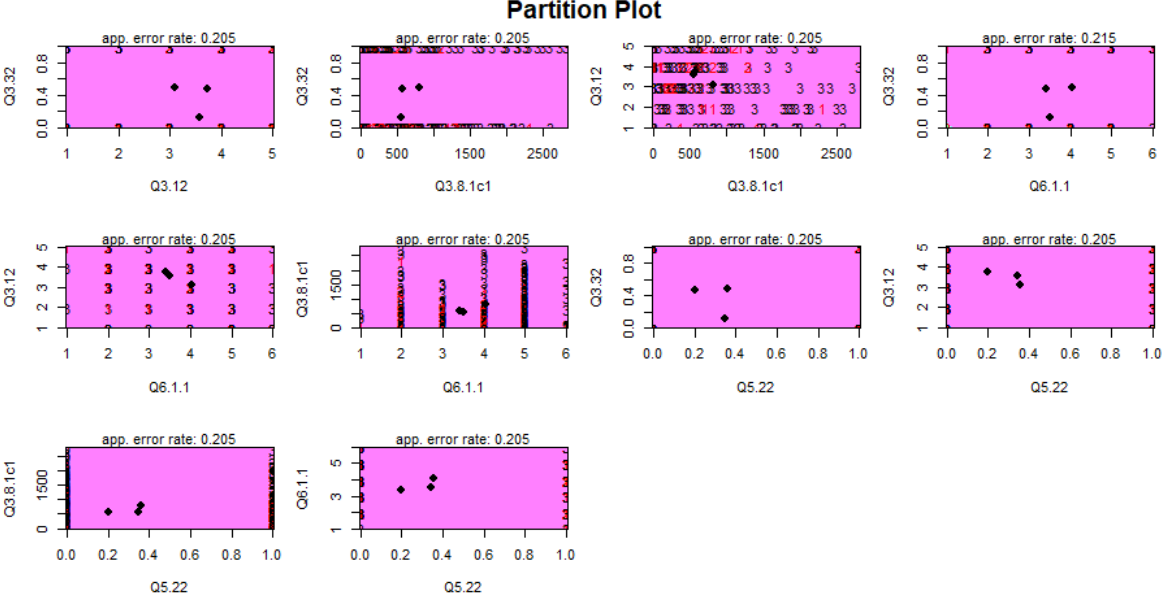


Figure 18 Partition plot LDA (Accommodation)

Figure 18 shows the three groups with overlapping boundaries. In this case, all the 3 Groups have overlapping boundaries and hence the model predicts all the students to be in group 3 “Satisfied” with the accommodation. It is eventually difficult to separate the classes and predict unless we have balanced data.

#### Predictions

This model fit the test data correctly and misclassified only 20% observations. But it doesn’t predict group 2 correctly 100% of times as there is a complete overlap between the adjacent classes, hence it is very difficult to predict if a student is “Neutral” in terms of satisfaction with the accommodation. The model accuracy is 80% on the test data.

### Quadratic Discriminant Analysis

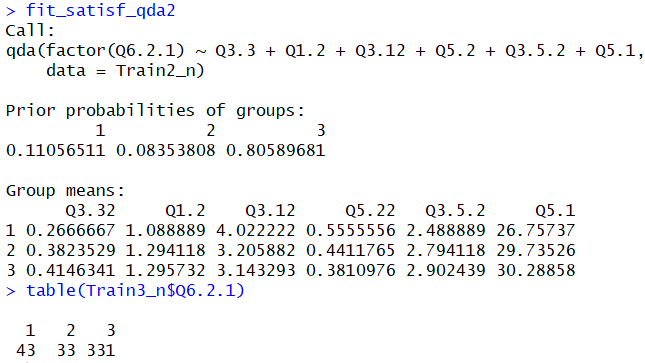
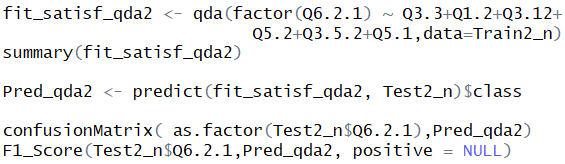
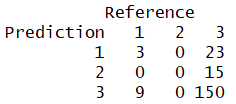
 

Figure 19 QDA (Accommodation)

#### Visualization

From Figure 20, we can visualize how the data among different classes are grouped.

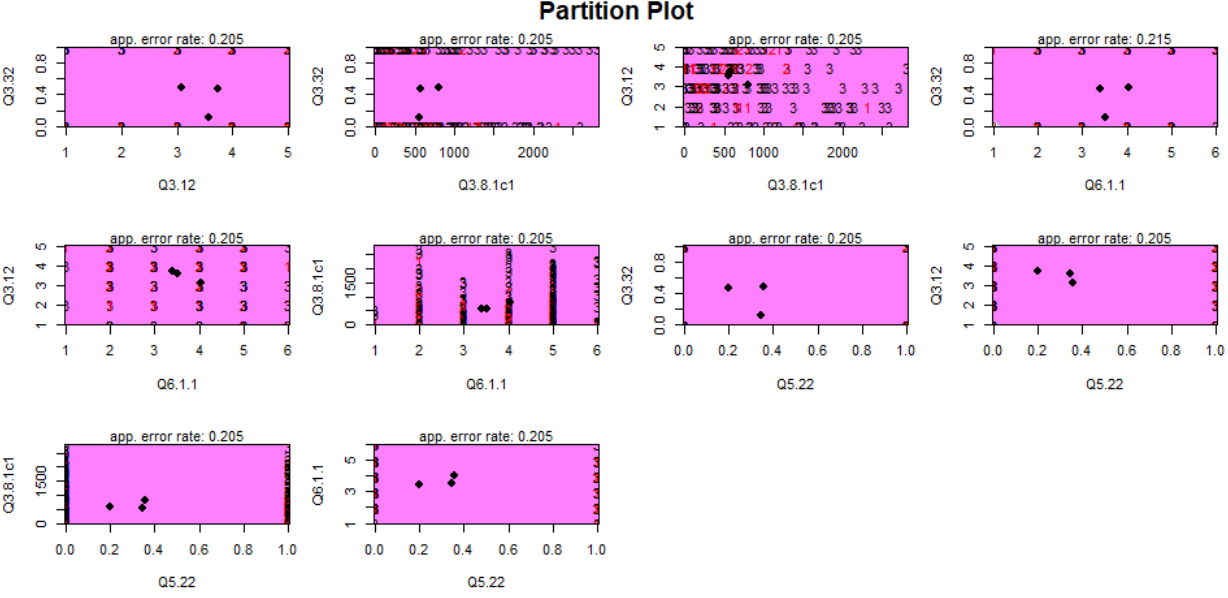


Figure 20 Partition plot QDA(Accommodation)

Again, the model predicts most of the students to be in group 3 “Satisfied” with the accommodation with indistinguishable boundaries with its neighbors.

#### Predictions

There are 407 observations, 43 in class 1, 33 in class 2 and 331 in class 3. QDA model assists in predicting the test data correctly but misclassified slightly more that LDA with 23% observations. Again Group 2 (“Neutral” in terms of satisfaction with the college) is incorrectly predicted 100% times. Overall it predicts student satisfaction accommodation with an accuracy of 77%.

## 4.5 Satisfaction with Financial/Material wellbeing Modeling

### Classical Linear Regression

Classical regression analysis helps to understand the relationship between dependent and independent factors. But in our case the data is not normally distributed and has a non-linear shape, thus findings of the analysis are indecisive and not used in results formulation and model comparison at a later stage.

Coefficients:

Estimate Std. Error t value PR(>|t|)

(Intercept) 3.048e+00 1.951e-01 15.619 <2e-16 \*\*\*

Q3.12 -4.564e-01 2.887e-02 -15.809 <2e-16 \*\*\*

Q6.1.1 7.781e-02 4.151e-02 1.875 0.0616.

Q5.1 -3.778e-03 4.474e-03 -0.845 0.3989

Q5.42 5.516e-02 8.602e-02 0.641 0.5217

Q3.7 3.630e-05 5.270e-05 0.689 0.4913

Q3.8.1c1 6.671e-05 6.058e-05 1.101 0.2715

Q5.22 -5.932e-02 6.956e-02 -0.853 0.3943

Q6.1.2 2.222e-02 4.120e-02 0.539 0.5900

Table 7 Linear Regression coefficients (Wellbeing)

|  |  |
| --- | --- |
| **Predictor** | **Significant** |
| Q1.12 - Entered college as a mature student? | **N** |
| Q5.4 - Would you consider yourself: Irish or Non-Irish | **N** |
| Q5.2 - e\_gender | **N** |
| Q7.5 - Social Class | **N** |
| Q3.7 - Total monthly disposable income | **N** |
| Q5.1 - Age (years)? | **N** |
| Q6.1.1 - Felt cheerful indicator (over last 2 weeks) | **Y** |
| Q3.12 - Experiencing financial difficulties indicator? | **Y** |

Table 8 Significance Evaluation (Wellbeing)

#### Interpretation

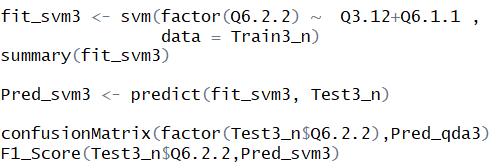
Students perceived satisfaction with financial/material wellbeing is derived from predictors **(P-Value<0.05)** using

Table 7.

* *Perceived level of financial difficulties* – Negative coefficient sign suggests that satisfaction level decreases for students facing higher levels of financial difficulty.
* *Feeling of being in good spirits and cheerful (over the last two weeks)* - Positive coefficient suggests that satisfaction level rises for students have felt more cheerful over last two weeks on a scale of 1-6 (At no Time - All the Time).

Again, some of the predictors which were assumed to be significant before the analysis on a general perception were found to be insignificant. Factors like **income, social class, Irish born, and age were marked as insignificant**; i.e. they don’t impact the overall student satisfaction with wellbeing. These predictors do not provide enough information to the model estimation and hence marked as insignificant while predicting the student’s satisfaction with financial/material wellbeing.

### Support Vector Machine



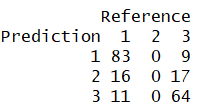
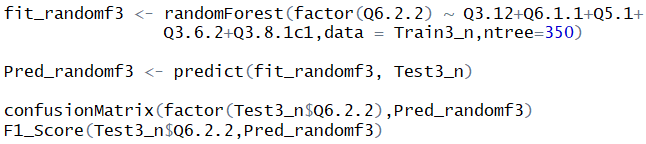


Figure 21 SVM (Wellbeing)

#### Interpretation

I have used the multi-category support vector machine algorithm fromthee1071 package (5). **SVM** fits the test data correctly and misclassified 26% observations. Again, in this case, Group 2 (“Neutral” in terms of satisfaction with wellbeing) is misclassified 100% of the time. Overall accuracy is 74% and it does a reasonable job in predicting “Unsatisfied” and “Satisfied” students.

### Random Forest



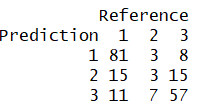


Figure 22 Random Forest (Wellbeing)

#### Interpretation

**Random Forest** fits the test data correctly and misclassified more than a quarter (27%) of observations. This time it does not completely misclassify all observations in Group 2 (“Neutral” satisfaction. Overall accuracy is 73% and it does a fair job in predicting all classes of students for satisfaction with wellbeing.

#### Visualization

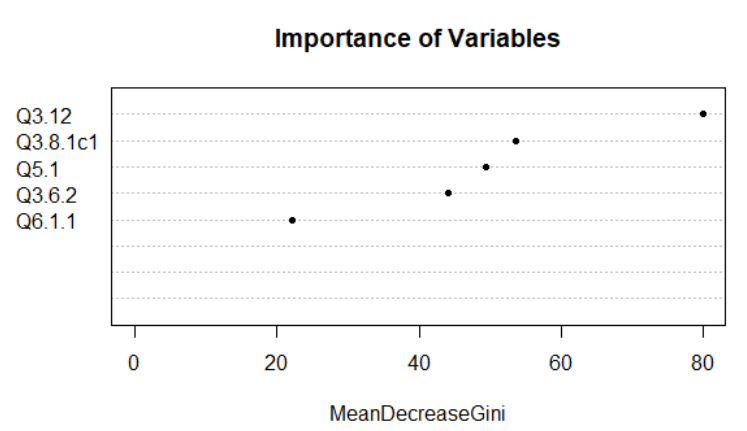


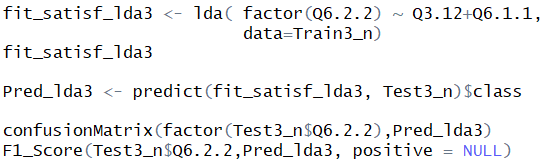
Figure 23 Importance Ranking (Wellbeing)

Figure 23 depicts “Importance ranking” for predictors: -

1. Q3.12 – Level of financial difficulties indicator?
2. Q3.8.1c1 - Total Expenditure - Own Pocket
3. Q5.1 - Age (years)?
4. Q3.6.2 - Distance between your house to the university
5. Q6.1.1 - Felt cheerful indicator (over last 2 weeks)

### Linear Discriminant Analysis

LDA has an assumption that there is a normal distribution in data among predictor space. As stated earlier that our data isn’t normally distributed, but we are still interested in the prediction accuracy, hence we implement LDA.



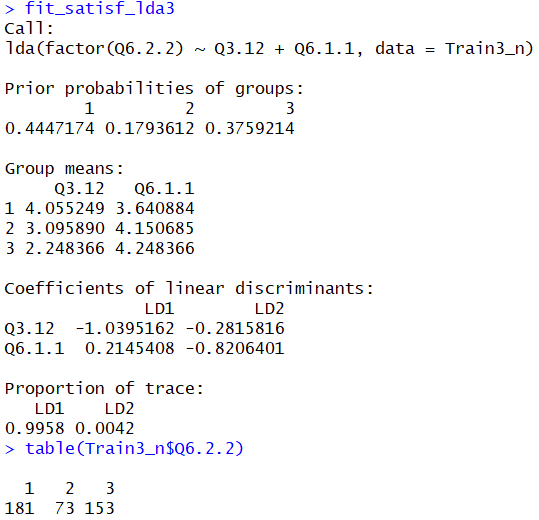
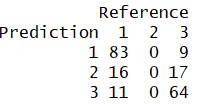
 

Figure LDA (Wellbeing)

#### Interpretation

There are 407 total observations, 181 in class 1, 73 in class 2 and 153 in class 3. [Figure 32]

LD1 is the first linear discriminant which is a linear combination of below parameters: -

**(Q3.12\*-1.0395162 + Q6.1.1\* 0.2145408)**

Q3.12 has a negative coefficient and Q6.1.1 has a positive coefficient, this is like the results obtained from linear regression. **LD1 explains almost all (99.58%) of the variability in the data.**

#### Visualization

From the Partition plot (Figure 25), we can visualize how the data among different classes are grouped together.

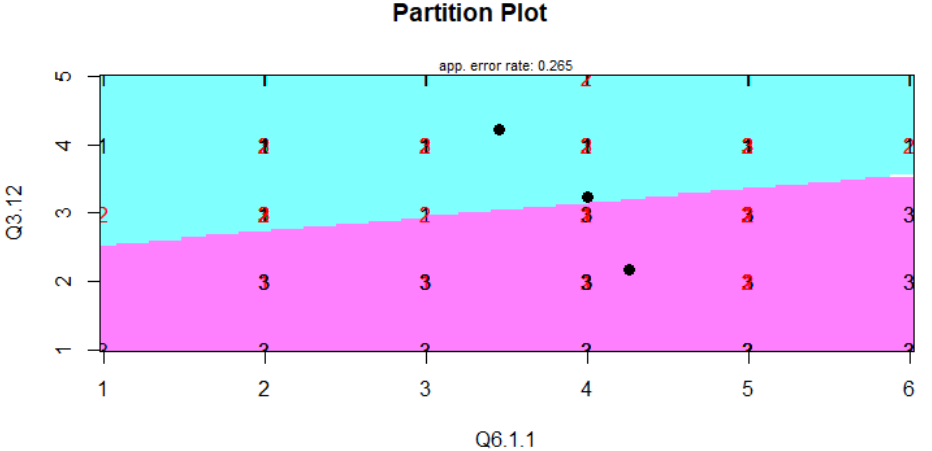


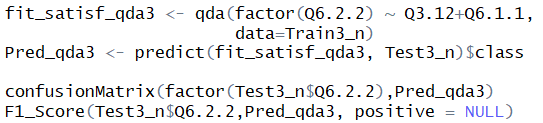
Figure 25 Partition plot (Wellbeing)

Figure 25 shows that Group 1 and 3 are distinguishable, but Group 2 has complete overlapping with neighboring boundaries and hence it is difficult to correctly predict if a student is “Neutral” in terms of satisfaction with the wellbeing.

#### Predictions

This model fit the test data correctly and misclassified 26% observations. But it doesn’t predict group 2 correctly 100% of times, as there is a complete overlap between the adjacent classes. The model has reasonably well accuracy on the test data of 74% for predicting of student satisfaction with the wellbeing.

### Quadratic Discriminant Analysis



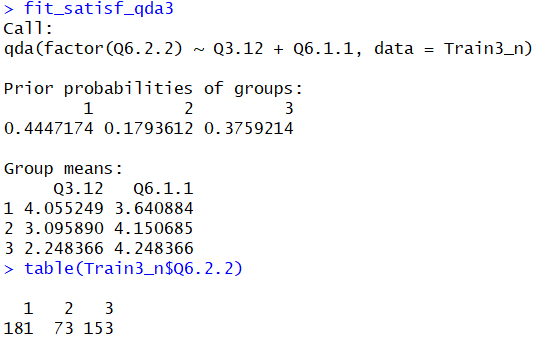
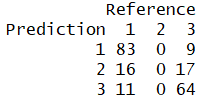
 

Figure 26 QDA (Wellbeing)

#### Visualization

Using the Partition plot for QDA, we can visualize how the data among different classes are grouped together.

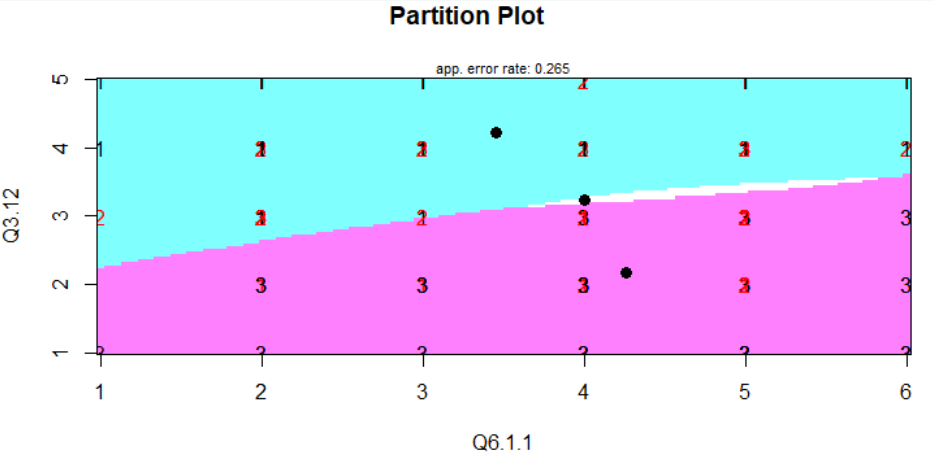


Figure 27 Partition plot (Wellbeing)

Figure 27 shows the shaded region depicting different classes like LDA, but with non-linear boundaries. The plot depicts all the students to be in either Group 1 “Unsatisfied” or Group 3 “Satisfied”, but Group 2 “Neutral” has an overlapping boundary with remaining groups and hence it is difficult to correctly predict if a student is “Neutral” in terms of satisfaction with the wellbeing.

#### Predictions

There are 451 observations, 43 in class 1, 42 in class 2 and 366 in class 3.QDA performs the same as LDA, in this case, misclassified 26% observations. Again Group 2 (“Neutral” in terms of satisfaction with the college) is incorrectly predicted 100% times. Overall it predicts student satisfaction with wellbeing with an accuracy of 74%.

# Chapter 5

## 5.1 Analysis of the predictor-response relationship

In this chapter, I have described the comparison of metrics from different ML algorithms implemented in the previous chapter. Multivariate modeling permits concurrent examination of the various predictors as well as deriving their relationship with different perceived satisfaction levels among students. The method considers, students’ satisfaction with below: -

* The College they are attending
* Accommodation
* Financial/Material wellbeing

Before moving to the actual comparison, let’s understand the metrics used to compare the results.

### 5.1.1 Metrics

Following metrics [6] are used in analysis and model comparison:

* ***Classes:*** Classes can be actual and predicted. The actual class is obtained from the real dataset used for model fit initially. Predicted classes are derived from the model itself using prediction techniques.
* ***Confusion Matrix*:** Comparison of the observed classes with the predicted classes.
* ***Sensitivity (true positive rate):*** The proportion of actual positives that are correctly identified from the final model. The more the better is the model fit.
* ***Specificity (true negative rate):*** The proportion of actual negatives correctly identified from the final model. It is critical in studies where a single wrong negative prediction can lead to huge losses.
* ***Type I Error (false positive rate):*** The cases where the predicted result is positive, but it is not.
* ***Type II Error (false negative):*** The instances when the predicted result is negative, but it is not.
* ***Accuracy:*** It is the description of systematic errors. The ratio of correct predictions made by the model to the total number of data points in the observation set.
* ***Precision:*** It is a description of random errors. The degree to which repeated measurement gives the same results under similar conditions. The ratio of truly predicted positive value to the total predicted positive values.
* **Recall:** Recall is the ratio of truly predicted positive values to all the data points that are positive.
* **F1 Score:**F1 score is the average of both precision and recall. The score closest to 1 are the best value closer to 0 are signs of incorrect model and prediction.

## 5.2 Satisfaction with the college

Students from 26 different institutes participated in the survey. Overall satisfaction with College among students is high where 80.7% of students are satisfied and only 9.1% of students are dissatisfied as shown in below Figure 28. 10.2% of students are neither satisfied or dissatisfied. [Figure 28]

Figure 28 Satisfaction with College

|  |  |  |  |
| --- | --- | --- | --- |
| **Technique** | **Accuracy** | **Misclassification** | **F1 Score** |
| **Support Vector Machine** | 0.78 | 0.22 | 0.44 |
| **Random Forest** | 0.77 | 0.23 | 0.5 |
| **Linear Discriminant Analysis** | 0.78 | 0.22 | 0.54 |
| **Quadratic Discriminant Analysis** | 0.78 | 0.22 | 0.49 |

Table 9 Comparison of Models (College)

Table 9 shows all the validated models and their performance with the given dataset.

When all the models are compared in terms of misclassification and F1 score, LDA has performed the best with misclassification of 0.22 and F1 score 54%. Thumb rule suggests that the model with an F1 score closest to 1 (100 % correct classification) is the best model.

Alternative options are Random Forest / QDA which have performed almost equally well with about 77 and 78% accuracy. We derive the final set of predictors from the best-selected model and now understand the relationship of the predictors with the satisfaction. First, we will discuss the nominal approach and then the ordinal approach.

### Multinomial Logistic Regression

Now once we have established the final set of predictors using the above chapter, we can analyze their relationship with the response.

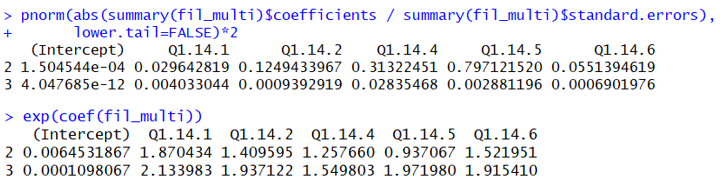
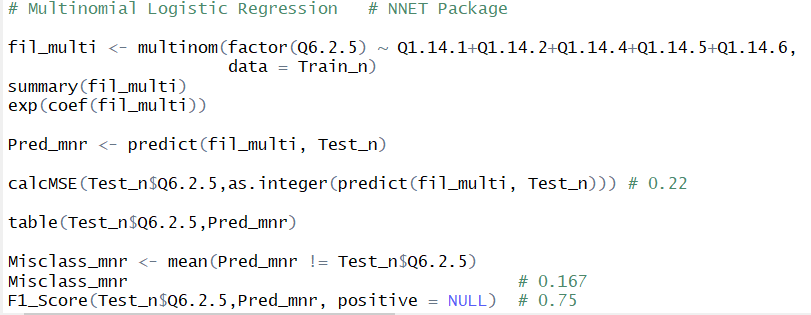


Figure 29 Multinomial Logistic Regression (College)

Two models are tested in this multinomial regression, one comparing Satisfaction of Level 1 to 2 and one comparing Satisfaction of Level 1 to 3 with the help of regression estimates as per below equations: -

Logit ([Neutral]/[Unsatisfied]) = β0+β1(Q1.14.1) + β2(Q1.14.2) + β3(Q1.14.4) + β4(Q1.14.5) + β5(Q1.14.6)

Logit ([Satisfied]/[Unsatisfied]) = β0`+β1` (Q1.14.1) + β2` (Q1.14.2) + β3` (Q1.14.4) + β4` (Q1.14.5) + β5` (Q1.14.6)

#### Interpretation

* **Neutral vs Unsatisfied Students** - With P-values (< 0.05) only Q1.14.1 is significant.

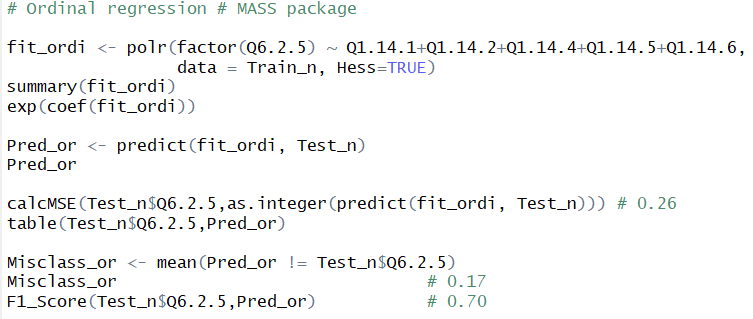
For a one-level increase in satisfaction for Quality of teaching, we expect to see an 87% (OR = 1.87) rise in the odds of being Neutral vs Unsatisfied in terms of satisfaction from college.

* **Satisfied vs Unsatisfied Students** - With P-values (< 0.05) all predictors Q1.14.1, Q1.14.2, Q1.14.4, Q1.14.5, Q1.14.6 are significant.

For example, we expect to see 113% (OR = 2.13) increase in the odds of being Satisfied vs Unsatisfied for one level increase in satisfaction of Quality of teaching. Similarly, odds of being Satisfied vs Unsatisfied for one level increase in satisfaction of Q1.14.2, Q1.14.4, Q1.14.5, Q1.14.6 increases by 93%, 54%, 97% and 91% (OR = 1.937122 / 1.549803 / 1.971980 / 1.915410).

### Ordinal Logistic Regression

After having considered the nominal approach, we will now discuss the ordinal approach that helps to answer another interesting question that how the satisfaction responses vary when ranked from 1-3.



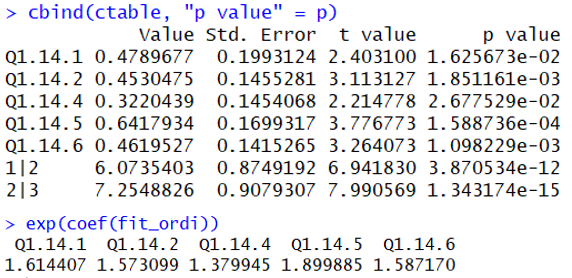


Figure 30 Ordinal Logistic Regression (College)

#### Interpretation

With P-values (< 0.05) all predictors Q1.14.1, Q1.14.2, Q1.14.4, Q1.14.5, Q1.14.6 are significant.

We expect to see 61% (OR = 1.61) increase in the odds of moving from Unsatisfied to Neutral or Satisfied for one level increase in satisfaction of Quality of teaching (Q1.14.1). Due to the property of **proportional odds assumption**, the same increase, 1.61 times, is found between Satisfied vs Neutral or Unsatisfied.

Similarly, odds of being Unsatisfied vs Neutral or Satisfied for one level increase in satisfaction of Q1.14.2, Q1.14.4, Q1.14.5, Q1.14.6 increases by 57%, 38%, 89.9% and 58.7% (OR =1.57 / 1.37 / 1.899 / 1.587). Similar conclusion for Satisfied vs Neutral or Unsatisfied category due to proportional odds assumption.

## 5.3 Satisfaction with the Accommodation

Postgraduate students belong to three main categories of accommodation — Landlord (41%), self/joint-owned (27%), or a parent’s property (20%). Students who were not paying for rent their accommodation belonged to highly satisfied (87%) category, compared to those who were renting (73%). For students on average, accommodation accounted for 38% of all expenditure (biggest single expense), accounting for around €420 per month. Overall satisfaction with Accommodation among students is high where 80.2% of students are satisfied and 11.7% of students are dissatisfied as shown in Figure 31. 8.1% of students are neutral.

Figure 31 Satisfaction with Accommodation

|  |  |  |  |
| --- | --- | --- | --- |
| **Technique** | **Accuracy** | **Misclassification** | **F1 Score** |
| **Support Vector Machine** | 0.8 | 0.2 | Na |
| **Random Forest** | 0.79 | 0.21 | 0.06 |
| **Linear Discriminant Analysis** | 0.8 | 0.2 | 0.07 |
| **Quadratic Discriminant Analysis** | 0.77 | 0.23 | 0.16 |

Table 10 Comparison of Models (Accommodation)

Table 10 shows all the validated models and their performance with the given dataset.

When all the models are compared in terms of misclassification and F1 score, QDA has performed the best with misclassification of 0.23 and F1 score 16%. Thumb rule suggests that the model with an F1 score closest to 1 (100 % correct classification) is the best model.

Alternate options are Random Forest / LDA for which the accuracy is 79% and 80%.

Significant (P-Value<0.05) predictors are; **gender,** **average monthly expenditure, financial difficulty indicator, felt cheerful indicator and rent paid indicator.** Using coefficients derived from Multinomial and Ordinal Regression, we can analyze the impact of significant predictors on the response variable. First, we will discuss the nominal approach and then the ordinal approach.

### Multinomial Logistic Regression

Now once we have established the final set of predictors using the above chapter, we can analyze their relationship with the response.

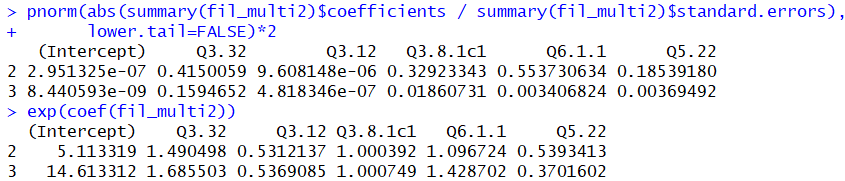
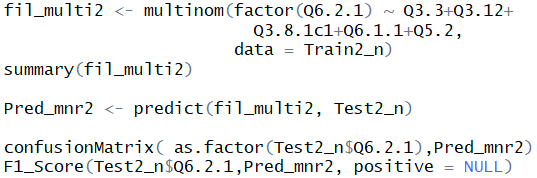


Figure 32 Multinomial Logistic Regression (Accommodation)

Using Figure 32, two models help to analyze two aspects using multinomial regression [Figure 32], one comparing Satisfaction of Level 1 to 2 and another comparing Satisfaction of Level 1 to 3 using regression estimates as per below equations: -

Logit ([Neutral]/[Unsatisfied]) = ß0+ß1(Q3.32) + ß2(Q3.12) + ß3(Q3.8.1c1) + ß4(Q6.1.1) + ß5(Q5.22)

Logit ([Satisfied]/[Unsatisfied]) = ß0`+ß1` (Q3.32) + ß2` (Q3.12) + ß3` (Q3.8.1c1) + ß4` (Q6.1.1) + ß5` (Q5.22)

#### Interpretation

**Neutral vs Unsatisfied Students** - With P-values (< 0.05) only Q3.12 is significant. For a one-level increase in perceived financial difficulties on a scale of 1-5 (No Difficulty to Serious Difficulties), we expect to see a **53%** (OR = 0.53) decrease in the odds of being Neutral vs Unsatisfied in terms of satisfaction with the accommodation. There exists an inverse relationship between satisfaction with accommodation and level of financial difficulty, which seems logical.

**Satisfied vs Unsatisfied Students** - With P-values (< 0.05) predictors Q5.22, Q6.1.1, Q3.12, and Q3.8.1c1 are significant. For understanding the relationships, we need to analyze the odds ratio for all the predictors as below: -

* Q3.12- We expect to see 53% (OR = 0.53) decrease in the odds of being Satisfied vs Unsatisfied for one level increase in perceived financial difficulties on a scale of 1-5.
* Q5.22 - For Male vs Female, we expect to see 37% (OR = 0.53) decrease in the odds of being Satisfied vs Unsatisfied with accommodation. Female students are more satisfied with their accommodation comparatively.
* Q6.1.1 – We expect to see 43% (OR = 1.43) increase in the odds of being Satisfied vs Unsatisfied for one level increase in perceived feeling more cheerful on a scale of 1-6 (At no Time - All the Time). Hence, cheerful students are almost twice as more satisfied with their accommodation compared to other students.
* Q3.8.1c1 – With the increase in average monthly expenditure (Own Pocket in Euros) on a nominal scale, we expect to see 0.1% (OR = 1.001) increase in the odds of being Satisfied vs Unsatisfied with accommodation. This may be since students with higher expenditure may be paying higher rents as well for better houses, in turn leading to more satisfaction with the accommodation.

### Ordinal Logistic Regression

After having considered the nominal approach, we will now discuss the ordinal approach that helps to answer another interesting question that how the satisfaction responses vary when ranked from 1-3.

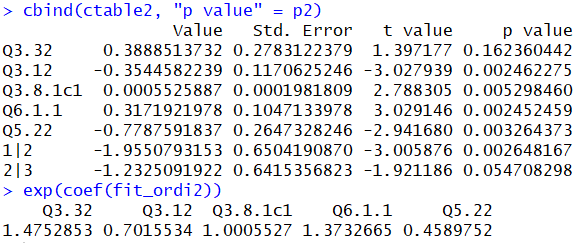
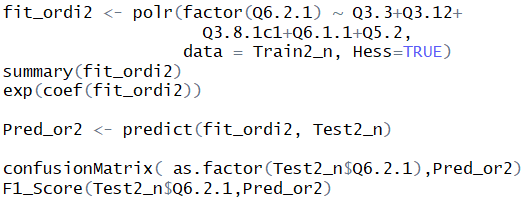


Figure 33 Ordinal Logistic Regression (Accommodation)

#### Interpretation

With P-values (< 0.05) predictors Q5.22, Q6.1.1, Q3.12 and Q3.8.1c1 are significant. [Figure 33].

We expect to see 70% (OR = 0.70) decrease in the odds of moving from Unsatisfied to Neutral or Satisfied for one level increase in perceived financial difficulties (Q3.12) on a scale of 1-5. Due to the property of **proportional odds assumption**, the same decrease, 0.7 times, is deduced between Satisfied vs Neutral or Unsatisfied.

Similarly, odds of being Unsatisfied vs Neutral or Satisfied for other predictors can be analyzed as below: -

* Q5.22 - For Male vs Female, we expect to see 46% (OR = 0.46) decrease. Again, this suggests that female counterparts are comparatively more satisfied with their accommodation.
* Q6.1.1 – We expect to see 37% (OR = 1.37) increase for one level increase in perceived feeling more cheerful on a scale of 1-6 (At no Time - All the Time). Hence, students who haven’t felt in good spirits lately are less satisfied with their accommodation.
* Q3.8.1c1 – With an increase in average monthly expenditure (*Own Pocket in Euros*) on a nominal scale, we expect to see a 0.1% (OR = 1.001) increase.

Because of the proportional odds assumption, the similar results apply for predicting Satisfied vs Neutral or Unsatisfied students.

## 5.4 Satisfaction with the Financial/Material wellbeing

The WHO-5 [16] score is presented as a measure of well-being. Figure 34 demonstrates that the major area of concern is the financial/material well-being of the surveyed students, where only 38% of all students were “satisfied”, on the other hand, 45% were dissatisfied.

Figure 34 Satisfaction with Financial/Material Wellbeing

|  |  |  |  |
| --- | --- | --- | --- |
| **Technique** | **Accuracy** | **Misclassification** | **F1 Score** |
| **Support Vector Machine** | 0.74 | 0.26 | 0.81 |
| **Random Forest** | 0.73 | 0.27 | 0.82 |
| **Linear Discriminant Analysis** | 0.74 | 0.26 | 0.82 |
| **Quadratic Discriminant Analysis** | 0.74 | 0.26 | 0.82 |

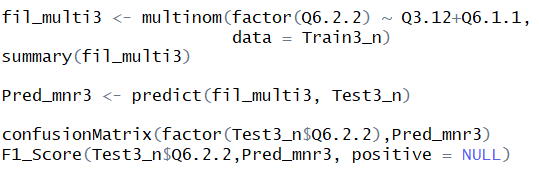
Table 11 Comparison of Models (Wellbeing)

Table 11 shows all the validated models and their performance with the given dataset.

When all the models are compared in terms of misclassification and F1 score, LDA and QDA have performed equally well with misclassification of 0.26 and F1 score 82%. Thumb rule suggests that the model with an F1 score closest to 1 (100 % correct classification) is the best model. Significant predictors are; **financial difficulty indicator felt cheerful indicator.** Using coefficients derived from Multinomial and Ordinal Regression, we can analyze the impact of significant predictors on the response variable. First, we will discuss the nominal approach and then the ordinal approach.

### Multinomial Logistic Regression

Now having established the final set of predictors in the above chapter, we can analyze their relationship with the response.



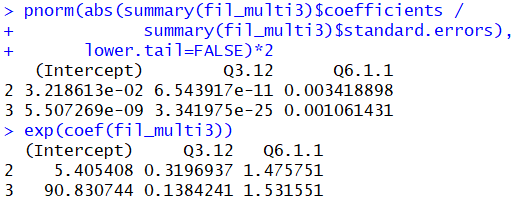


Figure 35 Multinomial Logistic Regression (Wellbeing)

Two equations are generated using multinomial regression [Figure 35], one comparing Satisfaction of Level 1 to 2 and another comparing Satisfaction of Level 1 to 3 for wellbeing using regression estimates as per the below equations: -

Logit ([Neutral]/[Unsatisfied]) = ß0+ß1(Q3.12) + ß2(Q6.1.1)

Logit ([Satisfied]/[Unsatisfied]) = ß0`+ß1` (Q3.12) + ß2` (Q6.1.1)

#### Interpretation

**Neutral vs Unsatisfied Students** - With P-values (< 0.05) significant parameters are.

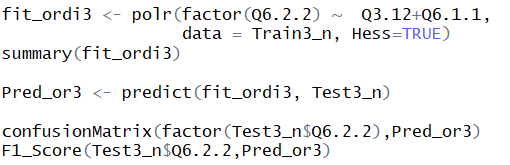
* Q3.12 - For a one-level increase in perceived financial difficulties on a scale of 1-5 (No - Serious Difficulties), we expect to see a 32% (OR = 0.32) decrease in the odds in terms of satisfaction with wellbeing.
* Q6.1.1 - We expect to see 48% (OR = 1.48) increase in the odds for one level increase in perceived feeling more cheerful on a scale of 1-6 (At no Time - All the Time).

**Satisfied vs Unsatisfied Students** - With P-values (< 0.05) significant parameters are.

* Q3.12- We expect to see a 14% (OR = 0.14) decrease in the odds for a level increase in perceived financial difficulties on a scale of 1-5. There exists an inverse relationship with financial difficulty and satisfaction with wellbeing.
* Q6.1.1 – We expect to see 53% (OR = 1.53) increase in the odds for one level increase in perceived feeling more cheerful on a scale of 1-6 (At no Time - All the Time). Hence, satisfaction with accommodation increases if the student has felt more cheerful over the last couple of weeks.

### Ordinal Logistic Regression

After having considered the nominal approach, we will now discuss the ordinal approach that helps to answer another interesting question that how the satisfaction responses vary when ranked from 1-3.



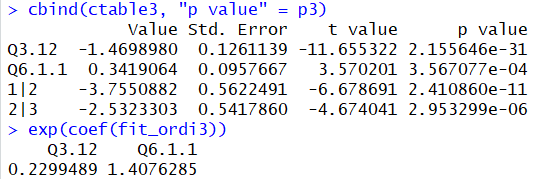


Figure 36 Ordinal Logistic Regression (Wellbeing)

#### Interpretation

With P-values (< 0.05) both the predictors are considered significant.

* Q3.12 - We expect to see 23% (OR = 0.23) decrease in the odds of moving from Unsatisfied to Neutral or Satisfied for one level increase in perceived financial difficulties on a scale of 1-5. There exists an inverse relationship with financial difficulty and satisfaction with wellbeing.
* Q6.1.1 – We expect to see 41% (OR = 1.41) increase in the odds of moving from Unsatisfied to Neutral or Satisfied for one level increase in the perceived feeling of being cheerful on a scale of 1-6 (At no Time - All the Time). Hence, satisfaction with accommodation increases if a student has felt more cheerful over the last couple of weeks.

Because of the proportional odds assumption, similar results apply for predicting Satisfied vs Neutral or Unsatisfied students.

# Chapter 6

## Conclusion

In the study, I found that certain algorithms can perform better with a given dataset. In this chapter, I have summarized the findings from different modeling techniques. I have also covered possibilities for future work.

The survey data for TMSI was filtered and analyzed to understand student’s perceived satisfaction with college, accommodation, and well-being. Significant predictors (based on P-value) were identified that influenced the response and the relationship was established with each response category. Following points are noteworthy: -

* Satisfaction is primarily a subjective concept. Student satisfaction is completely different across universities across the world and the indicators of different forms of satisfaction might vary when we move have different data under examination. Also, it is difficult to precisely identify each class of satisfaction uniquely as they are governed by factors that are overlapping in multidimensional space.
* LDA / QDA and Random Forest algorithm all performed very well (accuracy ~ 80%) with the classification problem and Logistic regression was remarkable in predicting the relationships between predictors and response.
* Teaching Quality, Organisation of timetable and studies, Teaching staff's attitude and facilities related to study are the most relevant predictors that students consider while rating their satisfaction with college. Satisfaction with accommodation varies with gender, average monthly expenditure, financial difficulty indicator, felt cheerful indicator and rent paid indicator. Financial difficulty indicators and felt cheerful indicator were the only important factors that governed well-being satisfaction.
* The analysis can answer questions like “How is satisfaction level Dissatisfied related to Neutral category” (Nominal approach) and “How is satisfaction level Dissatisfied related to Neutral or Satisfied category” (Ordinal approach).

## Future Work

This thesis provides important insights that may assist universities to analyze and improve the dimensions of the education system and administrative methods as suggested in the study by Sfakianaki et al. [25] Additionally, analysis outcome can be utilized by the Irish government to enhance the programs and institutions transitioning to the increased level of satisfaction among students. I have completed the proposed analysis under my thesis structure, still, I have highlighted possible future enhancements. I have summarised those as below: -

* Extend the analysis to all categories of students including Undergraduate and Ph.D. students.
* Comparison of the results from the Eurostudent IV survey data.
* Create a Shiny Application for Random Forest / LDA for predicting student satisfaction with tuning parameters variation capability.

# References

1. A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.
2. Athiyaman, A. (1997). Linking student satisfaction and service quality perceptions. The case of university education. European Journal of Marketing, 3(7/8), 528–540.
3. Cacoullos, T. (1973). Discriminant analysis and applications. New York: Acad. Press.
4. Carney, R. (1994). Building an image. Paper presented at the Proceedings symposium for the marketing of Higher Education, New Orleans, and Louisiana: American Marketing Association.
5. David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-1. https://CRAN.R-project.org/package=e1071
6. Domijan, K. (2018). ST464 Statistical machine learning.
7. Duque, L. C. (2014). A framework for analyzing higher education performance: students' satisfaction, perceived learning outcomes, and dropout intentions. Total Quality Management & Business Excellence, 25(1–2), 1– 21. <https://doi.org/10.1080/14783363.2013.807677>
8. Eurostudent V: Ireland 2013. Dublin: Higher Education Authority & Dublin: Insight Statistical Consulting [producers], 2013. Dublin: Irish Social Science Data Archive [distributor], 2014.
9. European Commission/EACEA/Eurydice, 2018. The European Higher Education Area in 2018: Bologna Process Implementation Report. Luxembourg: Publications Office of the European Union.
10. El Ansari W. Appraisal skills as a public health competency for evidence-based care: students' satisfaction with their research education. Journal of public health management and practice: JPHMP. 2004 Jul-Aug;10(4):354-65.
11. Finn, M. and Darmody, M. (2016). What predicts international higher education students’ satisfaction with their study in Ireland? Journal of Further and Higher Education, 41(4), pp.545-555.
12. Garrett, R. (2014). Explaining International Student Satisfaction: Insights from the International Student Barometer. Available online: <http://igraduate.dev.ingelby.com/assets/Explaining-Satisfaction-ISB-2014.pdf>
13. Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. <https://CRAN.R-project.org/package=tidyverse>
14. Hadley Wickham and Evan Miller (2019). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.1.0. <https://CRAN.R-project.org/package=haven>
15. HEA (2013) Higher Education Key Facts and Figures 2011/12 HEA: Dublin.
16. Heun R., Bonsignore M., Barkow K. and Jessen F. (2001) Validity of the Five Item WHO Well- being Index (WHO-5) in an Elderly Population European Archives of Psychiatry and Clinical Neuroscience 251(2).
17. Hosmer, D., Lemeshow, S. and Sturdivant, R. (2013). Applied Logistic Regression, 3rd Edition. 3rd ed. John Wiley & Sons.
18. Jokar, Saeid. (2018). Research Skills Education: Student Satisfaction.
19. Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1 – 26.
20. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.
21. Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2019). skimr: Compact and Flexible Summaries of Data. R package version 1.0.6. https://CRAN.R-project.org/package=skimr
22. News.idg.no. (2019). 60+ R resources to improve your data skills (- Software). [online] Available at: http://news.idg.no/cw/art.cfm?id=F66B12BB-D13E-94B0-DAA22F5AB01BEFE7
23. Onditi, E. O., &Wechuli, T.W. (2017). Service quality and student satisfaction in higher education institutions: A review of literature. International Journal of Scientific and Research Publications, 7(7), 328–335.
24. Osborne, J. (2013). Best practices in data cleaning. Thousand Oaks, Calif.: SAGE.
25. Sfakianaki, E., Matsiori, A., Giannias, D. A., & Sevdali, I. (2018). Educational leadership and total quality management: Investigating teacher leadership styles. International Journal of Management in Education,12(4), 375.
26. Shai Shalev-Shwartz and Shai Ben-David (2014). ‘Understanding Machine Learning: From theory to Algorithms’. Published in 2014 by Cambridge University press
27. Skrbinjek, V. & Dermol, V. Tert Educ Manag (2019) 25: 101. https://doi.org/10.1007/s11233-018-09018-5 Predicting students’ satisfaction using a decision tree.
28. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
29. Udacity. (2019). How to Choose Between Learning Python or R First | Udacity. [online] Available at https://blog.udacity.com/2015/01/python-vs-r-learn-first.html [Accessed 20 Jul. 2019].
30. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0
31. Weihs, C., Ligges, U., Luebke, K. and Raabe, N. (2005). klaR Analyzing German Business Cycles. In Baier, D., Decker, R. and Schmidt-Thieme, L. (eds.). Data Analysis and Decision Support, 335-343, Springer-Verlag, Berlin.
32. Weisberg, S. (2005). Applied linear regression. Hoboken: John Wiley & Sons, Inc., p.19.

# Appendix

The R Code of this complete project is available in following Link (GitHub):

<https://github.com/pratikkum/Satisfaction-analysis-and-prediction-of-Taught-Masters-Students-in-Ireland>