

2018

## Predicting Happiness - Comparison of Supervised Machine Learning Techniques Performance on a Multiclass Classification Problem

Dorota Nieciecka  
*Technological University Dublin*

Follow this and additional works at: <https://arrow.tudublin.ie/scschcomdis>

 Part of the [Computer Sciences Commons](#)

---

### Recommended Citation

Nieciecka, Dorota (2018). *Predicting happiness - comparison of supervised machine learning techniques performance on a multiclass classification problem*. Masters dissertation, DIT, 2018.

This Dissertation is brought to you for free and open access by the School of Computer Science at ARROW@TU Dublin. It has been accepted for inclusion in Dissertations by an authorized administrator of ARROW@TU Dublin. For more information, please contact [arrow.admin@tudublin.ie](mailto:arrow.admin@tudublin.ie), [aisling.coyne@tudublin.ie](mailto:aisling.coyne@tudublin.ie), [vera.kilshaw@tudublin.ie](mailto:vera.kilshaw@tudublin.ie).

# **Predicting happiness – Comparison of Supervised Machine Learning Techniques performance on a multiclass classification problem**



**Dorota Nieciecka**

A dissertation submitted in partial fulfilment of the requirements of  
Dublin Institute of Technology for the degree of  
M.Sc. in Computing (Data Analytics)

**May 2018**

## DECLARATION

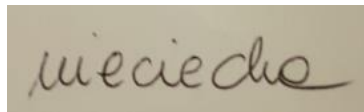
I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

**Signed:**

**Dorota Nieciecka**

A rectangular box containing a handwritten signature in dark ink, which appears to read 'nieciecka'.

**Date:**

**05 May 2018**

## ABSTRACT

In the modern world, especially in contemporary economies and politics, a population's subjective well-being is a frequent subject of the public debate. As comparisons of happiness levels in different countries are published, different circumstances and their effect on the value of the subjective well-being reported by people are also analysed. However, a significant amount of the research related to subjective well-being and its determinants is still based upon survey answers and employing conventional statistical methods providing details regarding correlations and causality between different factors and subjective well-being. Application of Supervised Machine Learning techniques for prediction of subjective well-being may provide new ways of understanding how individual factors contribute to the concept value and allow for addressing any issues, which may potentially affect mental and physical health.

The focus of this research is to use the survey data and make predictions regarding subjective well-being (a multiclass target) using Supervised Machine Learning models. In particular, the study is aimed at comparing the performance of two techniques: Decision Tree and Neural Networks. The 'C4.5 algorithm' used by the Decision Trees is considered as the benchmark algorithm, to which other supervised learning algorithms should be compared. At the same time, Neural Networks were previously proven to have high predictive power, even with multiclass categorisation problems.

Two experiments are conducted as part of this research, one using original highly imbalanced data; the other using the dataset balanced using SMOTE. The experimental results gathered show that for the first experiment there is no statistically significant difference ( $p < 0.01$ ) between models performance, while for the second experiment Neural Network's performance is lower than the one of a Decision Tree model with a statistically significant difference ( $p < 0.01$ ). With the 62.1% of the highest accuracy achieved, it is suggested that further research should be conducted to verify if any other Machine Learning model or approach to multiclass target classification could present better results when making prediction using survey data.

**Keywords:** Subjective Well-being, Supervised Machine Learning, Multiclass Classification, Imbalanced Data, SMOTE

## **ACKNOWLEDGEMENTS**

First and foremost, I would like to express my sincerest gratitude to my supervisor, Dr Deidre Lillis, for the patient guidance, constructive suggestions, expert advice and the encouragement provided throughout this research project. This accomplishment would not have been completed without her.

I would also like to thank all Dublin Institute of Technology (DIT) academic staff especially my professors at School of Computing for their knowledge, help, guidance and kindness, especially Dr Luca Longo, who has provided me with insight and the tools that I needed to choose the right direction and helped me at the start of this research project.

Additionally, I would like to acknowledge members of Healthy Ireland and ISSDA, who provided the resources used in this research (accessed via the Irish Social Science Data Archive - [www.ucd.ie/issda](http://www.ucd.ie/issda)), and declare that those, who carried out the original analysis and collection of the data, bear no responsibility for the further analysis or interpretation of it, done as a part of this research.

Finally, I would like to express my gratitude to my husband for providing me with continuous encouragement and support, not only through the process of researching and writing this thesis, but also throughout all years of study before it.

## TABLE OF CONTENTS

DECLARATION.....	I
ABSTRACT.....	II
ACKNOWLEDGEMENTS .....	III
TABLE OF CONTENTS .....	IV
LIST OF TABLES .....	VII
LIST OF FIGURES .....	VIII
1. INTRODUCTION .....	1
1.1. Background.....	1
1.2. Research Project .....	1
1.3. Research Objectives .....	2
1.4. Research Methodologies .....	3
1.5. Scope and Limitations.....	3
1.6. Document Outline .....	4
2. LITERATURE REVIEW AND RELATED WORK.....	6
2.1. Subjective well-being.....	7
2.1.1. Concept definition.....	7
2.1.2. Factors affecting SWB .....	8
2.1.3. Health Ireland Initiative .....	12
2.2. Machine Learning.....	13
2.2.1. Types of Machine Learning .....	14
2.2.2. Decision Trees .....	16
2.2.3. Neural Networks.....	18
2.2.4. Multiclass Classification Problem .....	18
2.3. Machine Learning in SWB Research.....	19
2.3.1. Gaps in Research .....	20
3. DESIGN AND METHODOLOGY .....	21
3.1. Business Understanding.....	23
3.2. Data Understanding.....	23

3.2.1.	Dataset Description.....	23
3.2.2.	Target Variable Investigation.....	24
3.2.3.	Feature Investigation.....	25
3.2.4.	Feature selection .....	25
<b>3.3.</b>	<b><i>Data preparation</i></b> .....	<b>26</b>
3.3.1.	Missing Values Handling .....	26
3.3.2.	Feature Selection.....	26
3.3.3.	Imbalance reduction.....	26
3.3.4.	Normalisation .....	28
<b>3.4.</b>	<b><i>Modelling</i></b> .....	<b>29</b>
<b>3.5.</b>	<b><i>Evaluation</i></b> .....	<b>29</b>
3.5.1.	Model performance comparison .....	30
3.5.2.	Statistical significance and hypothesis evaluation.....	33
<b>3.6.</b>	<b><i>Software</i></b> .....	<b>34</b>
<b>3.7.</b>	<b><i>Strength and limitations</i></b> .....	<b>35</b>
<b>4.</b>	<b>IMPLEMENTATION AND RESULTS</b> .....	<b>36</b>
<b>4.1.</b>	<b><i>Business understanding</i></b> .....	<b>36</b>
<b>4.2.</b>	<b><i>Data Understanding and Preparation</i></b> .....	<b>36</b>
4.2.1.	Target Variable .....	37
4.2.2.	Independent variables – missing values handling .....	38
4.2.3.	Independent variables – feature selection .....	40
4.2.4.	Feature investigation.....	42
4.2.5.	Data normalisation .....	45
4.2.6.	Target variable - imbalance removal .....	45
<b>4.3.</b>	<b><i>Modelling</i></b> .....	<b>46</b>
4.3.1.	Decision Tree modelling .....	49
4.3.2.	Neural Networks modelling .....	52
<b>4.4.</b>	<b><i>Evaluation</i></b> .....	<b>53</b>
4.4.1.	Model performance comparison .....	53
4.4.2.	Statistical significance and hypothesis evaluation.....	58
<b>4.5.</b>	<b><i>Experiment summary</i></b> .....	<b>61</b>

<b>5. ANALYSIS AND DISCUSSION.....</b>	<b>62</b>
5.1. <i>Strength and limitations of results .....</i>	<i>62</i>
5.2. <i>Considerations in regards to previous research .....</i>	<i>64</i>
<b>6. CONCLUSION.....</b>	<b>66</b>
6.1. <i>Research Overview .....</i>	<i>66</i>
6.2. <i>Problem Definition.....</i>	<i>66</i>
6.3. <i>Design/Experimentation, Evaluation &amp; Results.....</i>	<i>67</i>
6.4. <i>Contributions and impact.....</i>	<i>68</i>
6.5. <i>Future Work &amp; recommendations .....</i>	<i>69</i>
<b>BIBLIOGRAPHY .....</b>	<b>70</b>
<b>APPENDIX A: SAS CODE .....</b>	<b>79</b>



## LIST OF TABLES

Table 3.1 Target variable question and answers details.....	24
Table 3.2 Fit Statistics Grouping by Prediction Type.....	31
Table 4.1 Target variable mapping .....	37
Table 4.2 Variables rejected due to missing values count .....	39
Table 4.3 Selected Features Categorisation.....	41
Table 4.4 Feature selection algorithms output comparison.....	42
Table 4.5 Summary Statistics for selected features .....	43
Table 4.6 Count of individual classes before and after over-sampling.....	45
Table 4.7 Fit Statistics - models selected .....	53
Table 4.8 Model Performance - Misclassification Rates .....	55
Table 4.9 Wilcoxon Signed-Rank Test Results - Experiment 1 .....	58
Table 4.10 Wilcoxon Signed-Rank Test Results - Experiment 2.....	59

## LIST OF FIGURES

Figure 2.1 Overview of Literature Review and Related Work Chapter .....	6
Figure 2.2 Factors contributing to subjective well-being (Eurostat, 2005).....	11
Figure 2.3 Examples of Machine Learning Tasks (Kaplan, 2017) .....	16
Figure 2.4 Potential ID3-generated Decision Tree .....	17
Figure 2.5 Two-layer Neural Network (Han, Pei and Kamber, 2011).....	18
Figure 3.1 Phases of the CRISP-DM reference model (Chapman, et. al. 2000) .....	21
Figure 3.2 High-level Experiment Design Phases (by author) .....	22
Figure 3.3 Imbalance removal techniques (Pozzolo, 2016) .....	27
Figure 3.4 Schematic representation of 10-fold cross validation .....	30
Figure 3.5 Confusion Matrix .....	32
Figure 3.6 Example ROC curve (Bradley, 1997) .....	33
Figure 3.7 Example Agreement Plot (Giavarina, 2015) .....	34
Figure 4.1 Target variable classes' distribution.....	38
Figure 4.2 Feature distribution ranges .....	44
Figure 4.3 Imbalanced vs. balanced target .....	46
Figure 4.4 Experiment workflows .....	47
Figure 4.5 Distribution of target class in cross validation folds .....	48
Figure 4.6 Initial Decision Tree created.....	49
Figure 4.7 Sub-tree Assessment Plots - over-fitting model.....	50
Figure 4.8 Decision Tree – selected model settings.....	51
Figure 4.9 Sub-tree Assessment Plots - model selected.....	51
Figure 4.10 Iteration Plots - selected NN model .....	52
Figure 4.11 Misclassification Rates of Model.....	54
Figure 4.12 Confusion Matrices for Imbalanced Data Experiment .....	56
Figure 4.13 Confusion Matrices for Balanced Data Experiment .....	57
Figure 4.14 Agreement Plot - Experiment 1 .....	59
Figure 4.15 Agreement Plot - Experiment 2 .....	60

# 1. INTRODUCTION

## 1.1. *Background*

The concept of happiness, and how to achieve it, was of considered by the philosophers throughout the ages. However, as no formal definition was ever created, the philosophical concept of “happiness” was renamed by the psychologists pioneering its scientific study who proposed the term “subjective well-being” to be used as an alternative (SWB; Diener, 1984).

In contemporary economies and politics, a population's subjective well-being (SWB) takes a central place in the public debate, where comparison of happiness levels in different countries are performed and different circumstances are discussed in the context of their effect on the value of the subjective well-being reported by people.

Research performed to date proved that the SWB is in fact extremely complex and affected by a variety of different factors, including, but not limited to, socio-demographic or economic circumstances, social relationships, as well as, general health and health related habits, i.e.: diet, exercise and/or alcohol consumption.(Gerdtham and Johannesson, 2001; Adler, Dolan, and Kavetsos, 2017; Benjamin, *et.al.* 2014).

As the research continues, the exploration of the concept of using Machine Learning, or more specifically, testing the accuracy of making predictions regarding ‘subjective well-being’, is an interesting area for research. Especially considering that there are various examples of successful applications of Machine Learning classification techniques for predictions in areas such as marketing and financial services, retail, travel, healthcare, sociology, and most recently social media, already exist (Finlay, 2014).

## 1.2. *Research Project*

The main purpose of the research is to build and compare performance of two Supervised Machine Learning models: Decision Trees and Neural Networks for the multiclass classification of the subjective well-being response. The data set used for

the experiment contains information regarding respondents self-reported level of subjective well-being, which was collected as a part of Healthy Ireland Survey. The remainder of the survey questions will be considered as the independent feature variables, thus they will be used to make predictions regarding the target variable. Two predictive models will be built and then tested using two sets of data: the original survey dataset (with imbalanced distribution of classes within the target variable), and the balanced dataset (where the instances for the minority classes was increased using SMOTE). The comparison of the classifiers performance will be then conducted, using the Misclassification Rate data collected in the cross-validation process and applicable statistical tests.

The main goal is to test the performance of two classification algorithms for the prediction of subjective well-being, where not only multiclass classification must be performed, but also the target variable is imbalanced, and to confirm if the performance difference present between the models is statistically significant. Thus, the main research question of this project can be defined as:

*Which of the classifiers: Decision Trees or Neural Networks is more accurate in predicting subjective 'well-being' with the use of specified economic, social and health related factors?*

The following hypotheses are considered to allow for addressing above research question:

*H0: There is a statistically significant difference in the value of prediction accuracy of the subjective well –being between Neural Networks and Decision Trees with p-value <0.01*

*H1: There is no statistically significant difference in the value of prediction accuracy of the subjective well –being between Neural Networks and Decision Trees with p-value <0.01*

### **1.3. Research Objectives**

As the main goal of this research is to compare the performance of two classification algorithms on a multiclass classification problem and imbalanced data, the following objectives will have to be achieved in order to reach it: performed as mentioned below:

- Perform a literature review of the research conducted in relation to subjective well-being and machine learning, including any research where Machine Learning algorithms were implemented
- Perform initial data exploration followed by data cleaning and feature selection
- Perform detail exploration of selected independent variables and apply any required changes, e.g. normalisation.
- Build predictive models implementing Decision Tree and Neural Network algorithms.
- Apply both models created on two datasets: balanced and imbalanced
- Verify and compare models' performances using statistical tests and the Misclassification Rate obtained from k-fold cross validation process
- Evaluate the study and present findings, conclusions and recommendations for future work

#### ***1.4. Research Methodologies***

The focus of the research is the comparison of performance of two well-known supervised machine learning models for prediction of self-reported level of subjective well-being obtained as a part of the Healthy Ireland Survey, thus it is considered as secondary research.

Additionally, in order to accept or reject the research hypotheses, secondary quantitative data will be used to conduct an experiment involving development, employment, and evaluation of Machine Learning models. Results obtained through this process will then be compared using appropriate statistical tests. Therefore, the research methods for the report can be summarized as secondary, quantitative, empirical, and deductive.

#### ***1.5. Scope and Limitations***

The major limitation of the research is the presence of imbalance in the multiclass target and fairly low correlation between the feature and the target variables, which may result in the low prediction accuracy of both models tested. Therefore, the two

models built will be tested on two datasets: the original survey dataset (with imbalanced distribution of classes within target variable), and the balanced dataset (where the instances for the minority classes was increased using SMOTE).

The first part of the experiment will allow for obtaining an insight on usefulness of the survey data in prediction of subjective feeling of happiness. The second part of experiment, where the balanced dataset will be used, will then provide the insight on how minority class imbalance removal may affect overall results. It is possible that the use of Synthetic Minority Over-sampling Technique on minority class may produce instances different than the real world data, which in effect may skew model results.

## **1.6. Document Outline**

This research report consists six chapters in total, except for Chapter 1, the following sections can be identified:

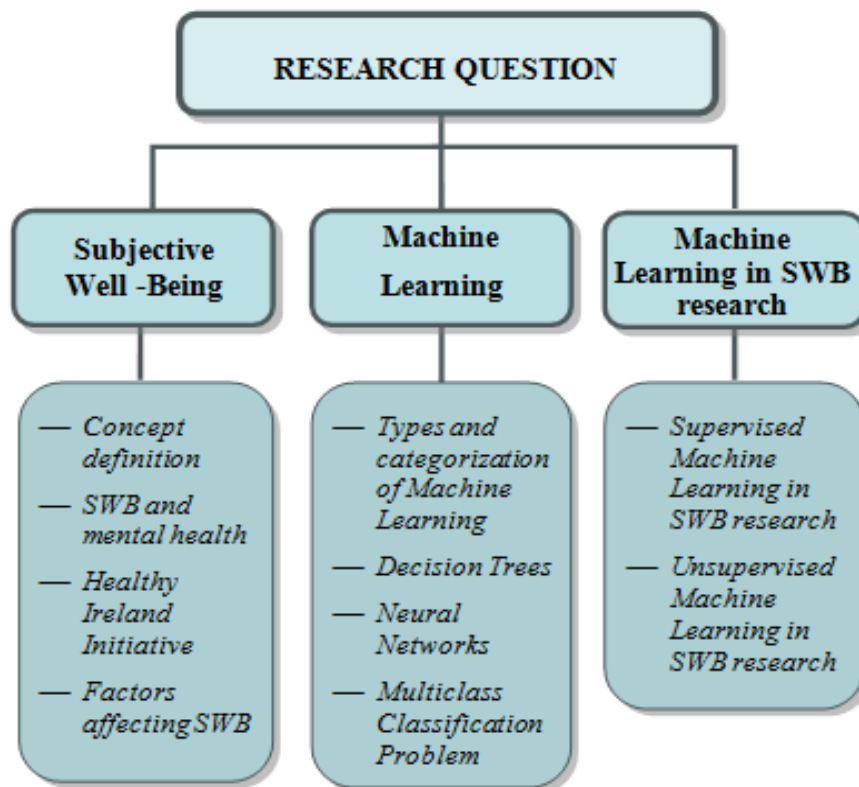
- **Chapter 2 - Literature Review and Related Work:** provides a review of research related to and important in terms of the research question and experiment. It is divided into three sections. The first one describes the concept of Subjective Well-Being, including its definition and affecting factors identified to date. Additionally, an overview of the Healthy Ireland Initiative is performed, to present not only the dataset background, but also the role of the Healthy Ireland Survey data collection in the process of improvement of overall well-being of the Irish population. The second part presents selected concepts related to Machine Learning, which were important for the decisions made in relation to experiment design and methodology. Finally, the last section presents SWB research to date, where Machine Learning was implemented as a tool, which serves as the foundation for identifications of any gaps present.
- **Chapter 3 – Design and methodology:** describes in detail the design of the experiment performed as a part of the research, including explanation of each phase and step, including the details regarding the software used for it.
- **Chapter 4 – Implementation and results:** describes the implementation of the experiment. It describes all the activities performed together with the actual results obtained; those include the preparation of the dataset, feature selection, model

building and adjustment, as well as model comparison in the cross-validation. The hypothesis testing is also performed here.

- **Chapter 5 – Analysis and discussion:** presents detailed analysis and evaluation of experimental results. The strength and weaknesses of the experiment are analysed here. Additionally, the findings obtained in the experiment are then compared to other findings that were previously identified and discussed in the literature review.
- **Chapter 6 – Conclusion:** provides a summary of the research undertaken including problem definition, critical analysis of the experiment design and implementation, as well as evaluation of the results. Additionally, it presents the discussion regarding possible improvements and future work.

## 2. LITERATURE REVIEW AND RELATED WORK

This chapter provides a review of the literature of research and the work related to the concept of subjective well-being and some concepts from machine learning, which will allow for answering the research question. Four areas discussed here are: Healthy Ireland Initiative, Subjective Well-Being, Machine Learning, and Application of Machine Learning in SWB Research, all of these are presented in the Figure 2.1 below.



**Figure 2.1 Overview of Literature Review and Related Work Chapter**

The first section of literature review is concerned with the concept of subjective well-being and its first emergence in the literature. The details of the concept will be presented together with its importance for one's overall mental health. The Healthy Ireland Initiative will also be discussed here, as the data used for the research was originally collected in a survey as a part of this initiative. Findings from the original report regarding survey data and other publications related to the subjective well-being will be reviewed, as they are of relevance to the research question. Analysis of research related to different groups of factors affecting SWB will be performed, which



will be followed by the comparison of groups of factors available in the dataset used for the experiment.

The second section describes some of the Machine Learning definitions and concepts which are relevant to the experiment performed in this research, including a review of previous research leading to approach selection, together with categorisation and definition of algorithms and strategies selected.

Finally, the last section of the review is related to the analysis of the research and publications related to the use of Machine Learning for the prediction or exploration of SWB, together with a discussion regarding its advantages and disadvantages.

The chapter is concluded with an analysis of gaps identified in all review sections leading to the research question definition. Brief discussion regarding motivation behind the research and the limitations is also made here.

## ***2.1. Subjective well-being***

This section describes the concept of Subjective Well-Being, including its definition and affecting factors identified to date. Additionally, an overview of the Healthy Ireland Initiative is performed, to present not only the dataset background, but also the role of the Healthy Ireland Survey data collection in the process of improvement of the overall well-being of Irish population.

### **2.1.1. Concept definition**

One of the definitions of the concept of subjective well-being (SWB) describes it as personal, emotional, and cognitive evaluation of individual's life. The name 'subjective well-being' is also alternatively called: happiness, peace, fulfilment, and life satisfaction (Diener, Oishi & Lucas, 2002).

Subjective well-being, or rather "happiness," has been of significant interest throughout most of human history. Unfortunately, there is no uniform definition of the concept and how it can be achieved. In fact, the definition has been debated for as long as philosophers have been inquiring the concept. Starting with Ancient Greece, an exploration of the nature of happiness was made by Democritus (460 BC–370 BC),

which was then followed by such philosophers as Socrates, Plato, or Aristotle , and many others throughout different eras, from the Middle Ages through to the Age of Enlightenment up to 19<sup>th</sup> century's Utilitarianism (Tatarkiewicz, 1976). Review of those provides proof that although different philosophies through centuries differ significantly from one another, most of the philosophers agreed on one thing, which is the difficulty of defining the happiness. Thus, the philosophical concept of “happiness” was renamed by the psychologists pioneering its scientific study; they proposed the term “subjective well-being” to be used as an alternative (SWB; Diener, 1984).

As previously mentioned, “subjective well-being” can be defined as individual's personal evaluation of their life. Thus, it includes both the cognitive judgment of life satisfaction and the appraisal of emotions. This definition of SWB emphasizes the subjective nature of the concept. (Diener & Suh, 1997). Nevertheless, although assessment of SWB is subjective by nature, the review of literature also provides the evidence that subjective well-being is affected by a number of separable although related factors. Thus, in order to understand the SWB, many researchers attempted to determine how individual components are affecting it. These are discussed in detail in section 2.1.2 of the research paper.

However, the purpose of this research is not to define the concept of SWB, but to verify the use of Machine Learning algorithms in the prediction of SWB using survey data, which includes questions related to different groups of factors. An ability to make predictions regarding one's SWB could be valuable, for example in relation to the WHO report: ‘Promoting Mental Health’, where an emphasis is made on SWB importance to overall mental health and possible negative outcomes resulting from low or negative subjective well-being (e.g. suicide, health deterioration, etc.) (WHO, 2005).

#### 2.1.2. Factors affecting SWB

Modern research, including studies conducted by psychologists, sociologists and economists, increased the understanding of how the individual components (or factors) affect the subjective well-being. The main groups of factors include: economic circumstances, social relationships, as well as health and health related habits. These are discussed in more detail below:

### — Economic Factors

The most frequently analysed economic factors are employment and income. Low income is always correlated with low SWB (Becchetti and Rossetti, 2009). The analysis of the literature in the field provided evidence that, while higher absolute income increases SWB, this positive correlation is present only up to a certain level (Mentzakis and Moro, 2009). The same findings are made in the research by Ferrer-I-Carbonell (2005), Pedersen and Schmidt (2011) and Frey and Stutzer (2005), where positive correlation between income and life satisfaction is present only until the presence of “frustrated achievement”, where the increase of income is in fact associated with the reduction in life satisfaction, related to a decrease in areas such as health and quality of social relationships.

Unemployment is the other factor of interest. Data analysis and results obtained by Gerlach and Stephan (1996) provide evidence of a consistent strong negative correlation between the SWB and unemployment, however with the different levels of SWB values between men and women. At the same time, Dolan, Peasgood, and White (2008), while also providing proof of overall lower levels of SWB for unemployed individuals, highlighted the importance of factors from other groups, namely the social relationship group.

Unfortunately, only employment status data, but not the income, is collected as a part of Healthy Ireland Survey used for this research. Thus, while it will be possible to analyse the correlation and its strength between the unemployment and the SWB, it will not be possible to identify any cases of “frustrated achievement”. As all the previous research discussed above showed strong negative correlation between the unemployment and the SWB, this variable may also be relevant to the predictive models.

### — Social Relationships

Social relationships are another group of factors, which were proved to have a strong influence on the subjective well-being Fernández-Ballesteros, *et.al.* (2001). The research conducted in relation to this group includes such individual factors as: family relationships (North *et.al.*, 2008), marriage (Schoon, Hansson and Salmela-Aro, 2005), and/or lack of social interaction (Umberson and Montez, 2010).

North *et.al.* (2008) examined the role of family life and its influence on the reported level of happiness. The authors proved that income had a small, positive impact on happiness, in contrast to family and social support, which had a strong positive relationship to change in happiness. Similar findings are also identified by Schoon, Hansson and Salmela-Aro (2005), who proved a positive correlation between successful marriage and overall feeling of happiness.

Alternatively, Umberson and Montez (2010) emphasize the importance of social relationships on the subjective well-being and overall health status resulting from it. An observation is made here that quantity and quality of social relationships (friendships, marriage, belonging to religious organization) have both short-term and long-term effects on life-satisfaction level.

As the data used for this research includes a set of question regarding social relationships, including marital status as well as social connectedness questions, it is possible that some of those variables will be selected in the feature selection process. Unfortunately, as no income data is present it won't be possible to compare the findings to the ones from the research by North, *et.al* (2008), however, the analysis of correlation will provide detail regarding, which of the two factors, social relationships or employment, are more significant for the prediction of SWB.

#### — **Health and Health Related Habits**

Finally, the last group of factors, which is frequently investigated in the SWB research, is the health and health related habits. Dolan, Peasgood, and White (2008) argues that there is a strong relationship between SWB and both physical and mental health, with the stronger correlation being present for the mental health than physical health. Nevertheless, it is proved that some specific conditions, such as heart attacks and/or stroke always negatively affects subjective well-being (Shields & Wheatley Price, 2005), with the causality being from the health condition to SWB. Oswald and Powdthavee (2006) present evidence between prolonged sickness and/or disability and the reduction of SWB.

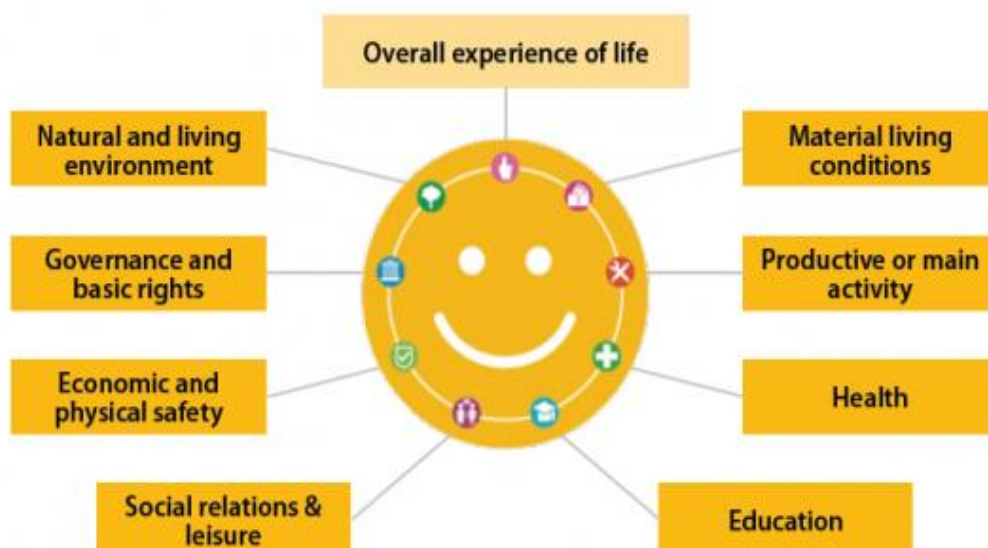
Alternatively, research by Stoica (2015) examines SWB in the different aspect and provides evidence of importance of sleep in self-assessed subjective well-being. It also shows that the happiness is not only a product of external factors, but is also based in part on circadian rhythms of an individual. Moreover, Coyle and Vera (2013) provide

evidence of strong negative correlation identified between unexpected or prolonged stress and the overall perceived SWB.

In regards to health related habits, research by Fox (1999) provided evidence that even simple types of exercise are associated with increased SWB, especially for individuals over 60 years old. Additionally, physical activity among those over 60 years old was also negatively associated with depressive symptoms (Baker et al., 2005).

Health and health related factors are the main focus of the Healthy Ireland Survey questionnaire. Thus, it is possible that these factors will dominate over other inputs used by the predictive models. Thus, it will be important to analyse, which of the health related factors is the most strongly correlated with the SWB value prediction.

All the above findings can be additionally considered in the context of government publications. The Eurostat reports (2015 and 2017) describe subjective well-being as a multidimensional concept of ‘overall experience of life’, which can be considered the key indicators in the Quality of Life determination (see Figure 2.2, source: Eurostat, 2005). Therefore, the analysis of factors selected for the purposes of building the predictive models will be made with consideration to the categorisation made in those reports.



**Figure 2.2 Factors contributing to subjective well-being (Eurostat, 2005)**

### 2.1.3. Health Ireland Initiative

Healthy Ireland is an initiative of the Irish Government with the purpose of improving the overall wellbeing of society, in terms of both physical and mental health. The initiative wants to reduce the risks to poorer health and wellbeing such as obesity, mental health problems, smoking, or alcohol abuse. Thus, one of the key activities of Healthy Ireland is to collect information not only about the health status, but also on how to improve it. This analysis also includes the collection of details related to the subjective well-being (Ipsos, 2016).

The data, which provides up-to-date information regarding the nation's health, is collected in the annual Healthy Ireland Survey, which is carried out by Ipsos MRBI on behalf of the Department of Health. The first survey was conducted between years 2014/2015 with the report on findings published in October 2015. The second survey was carried out between the years 2015/16 and a report of its findings published in October 2016<sup>1</sup>. It is the data from this survey that will be used for the purposes of this research (Department of Health, 2016).

The report on key findings published by Ipsos (2016) didn't include the analysis of the individual questions, but focused on key fact from different sections of survey and their correlation with one another. Thus, it was reported that overall positive mental health is more likely higher among men than women. Additionally, physical activity and financial stability were the changes most frequently selected by the respondents, as the ones which would improve their health and wellbeing (Ipsos, 2016).

Literature review performed in regards to the research on subjective well-being provided the evidence that subjective well-being is affected by a number of separable although related factors. The main groups of factors include: economic circumstances, social relationships, and health and health related habits. As the purpose of this research is to verify the use of Machine Learning algorithms in the prediction of SWB using survey data, part of the data pre-processing will be to select individual features,

---

<sup>1</sup> <http://www.healthyireland.ie/accessibility/healthy-ireland-survey/>

which includes questions related to above groups of factors. In effect it will be possible to categorise the factors selected, and compare their correlation to the target variable, with the one identified in the previous research.

## **2.2. Machine Learning**

This section presents selected concepts related to Machine Learning, which were important for the decisions made in relation to experiment design and methodology.

Machine learning is an interdisciplinary subfield in computer science that involves automated formulation of complex predictive models and algorithms through the use of multiple techniques from fields such as statistics, game theory, information theory and optimization. (Shalev-Shwartz and Ben-David, 2014). It was also defined as the process of converting experience into expertise (Carbonell, Michalski and Mitchell, 1983).

Machine Learning significantly evolved since its first emergence in Arthur Samuel's Checkers-playing program developed in 1952, and the initial work by Hunt, *et al* (1966) in inductive problem solving, Nilsson (1965) in statistical functions and data classification, Rosenblatt (1961) in Neural Networks, and Vapnik (1963) in Support Vector Machine. Currently, it is a widely acknowledged solution, which is used in such areas as pattern recognition, new knowledge development, and predictive analytics. (Siegel, 2016)

A review of literature related to Machine Learning provided a list of various examples of successful application of Machine Learning classification techniques for predictions making, regarding individuals in a large population, in areas such as: marketing and financial services, retail, travel, healthcare, sociology, and most recently social media (Finlay, 2014). This suggests that classification techniques could also be used for prediction making of the psychological concept of 'subjective well-being'

Additionally, as multiple different models can be listed, with the variety of them finding application in the commercial and public institutions, a review of empirical studies regarding model performance was performed. The findings from studies conducted by Caruana & Niculescu-Mizil (2006), Chavan, *et.al.* (2014), Zhang & Lee,

(2003) , (Iniesta, Stahl and McGuffin, 2016) and (Zoonen and Toni, 2016), which provide detailed comparison of various learning algorithms in different scenarios, including multiclass classification problems, allowed for the selection of Decision Trees and Neural Networks, as the algorithms which will be assessed to determine their effectiveness and accuracy in relation to prediction of ‘subjective well-being’. As the accuracy of the classifier always varies for each individual dataset (Moran, He, & Liu, 2009), an experiment was conducted to determine which of the classifiers, Decision Trees or Neural Networks, is more accurate in predicting subjective ‘well-being’ with the use of specified economic, social and health related factors.

Decision Trees were selected as they are the most fundamental machine learning models, which are able to provide interpretability and information about the importance of individual features. At the same time, Neural Networks were proven to outperform other models in multiclass categorisation when extension from binary is used (Pal and Mitra, 1992). Subsections 2.3.1 to 2.3.3 of this chapter will provide an overview of Machine Learning techniques chosen for the experimental purposes, as well as the main concept of interest related to them.

### 2.2.1. Types of Machine Learning

Machine Learning is a wide field, in which all the learning paradigms can be differently categorised. According to Shalev-Shwartz and Ben-David (2014), the following taxonomies should be considered:

#### — *Supervised versus Unsupervised*

The first division of Machine Learning algorithms is based on the nature of their interaction between the learner and the environment.

In Supervised Machine Learning (Shalev-Shwartz and Ben-David, 2014) an algorithm is presented with a set of input variables (X) and an output variable (Y), and the goal of creating the mapping function, which then can be used to predict the output variables (Y) from new set of input data. Thus, the process is called supervised because algorithm training involves the oversight over the prediction made and knowledge of the correct answer. Therefore, learning can be stopped when the algorithm performance becomes acceptable.



On the other hand, Unsupervised Machine Learning (Shalev-Shwartz and Ben-David, 2014) involves a learning process where an algorithm is presented only with the set of inputs (X) and no output variables are specified. The purpose here is to model the underlying data structure or its distribution in order to obtain insight regarding any hidden patterns. Thus, as no correct answer exist, it is impossible to have an oversight over the model and its performance.

#### — *Active versus Passive Learners*

Learning paradigms can also be categorised based on the role played by the algorithm, also called a learner, in addition to its interaction with the learning environment described in the previous point. It is possible to distinguish two types of learners: active and passive (Shalev-Shwartz and Ben-David, 2014). An active learner is a type which interacts with the environment during the training process (via queries or experiments), while a passive learner only uses the data provided without influencing it.

#### — *Online versus Batch Learning*

Another parameter to consider is the distinction between situations in which the learner is presented with the data in sequential order, known as online learning, as opposed to batch learning, where the algorithms is trained and generates the best predictor by learning on the entire training data set at once(Shalev-Shwartz and Ben-David, 2014).

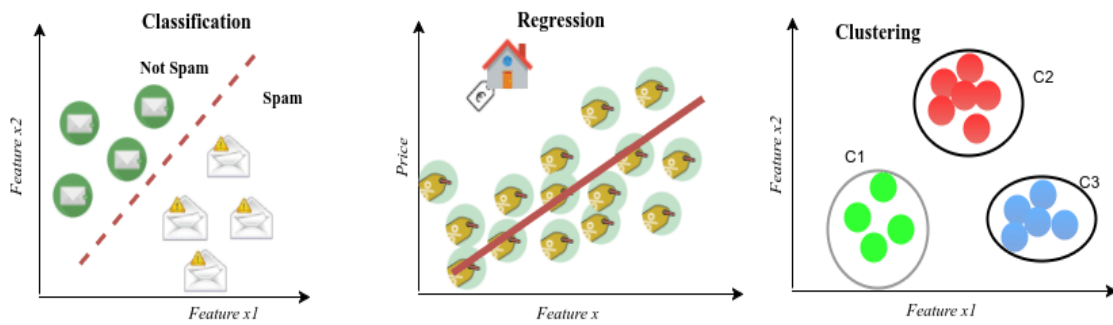
#### — *Learning Problem Type*

Finally, the last categorisation of Machine Learning algorithms is considered with the output that's being produced by the model (Shalev-Shwartz and Ben-David, 2014). The most important types include:

- **Classification**, where a model, also called classifier, is trained to identify the discrete class of a target variable, when a set of inputs is given. It is possible to distinguish here: *binary classification* (target variable has only 2 classes) or *multiclass classification* (where the problem can belong to one of three or more classes)
- **Regression**, where the outputs are continuous

- **Clustering**, where all inputs are being divided into groups. However, the groups are not known before the training of an algorithm starts (Shalev-Shwartz and Ben-David, 2014).

This research considers only a subset of presented learning paradigms. The main focus of the experiment performed is to analyse and compare the performance of supervised classification batch learning with both passive learner (Decision Tree) and active learner (Neural Network).



**Figure 2.3 Examples of Machine Learning Tasks (Kaplan, 2017)**

### 2.2.2. Decision Trees

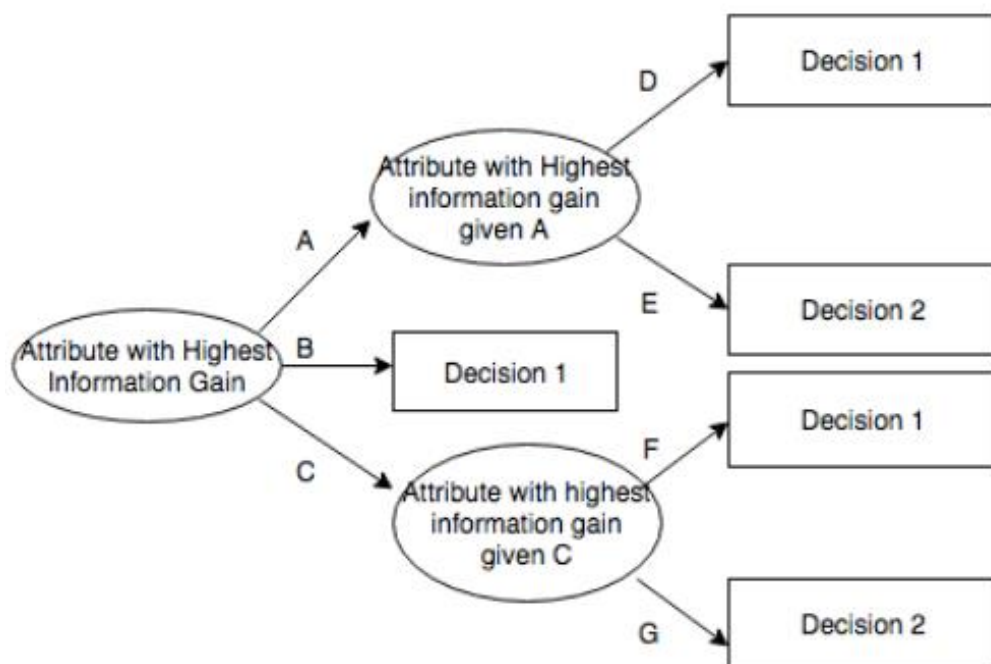
A decision tree models can be represented in a is a tree-like chart structure, where each internal node represents a test on an attribute, each branch denotes an outcome of the test, and each leaf node holds an output value. The top node in a tree is called the root node (Swain and Hauska, 1977).

As the construction of a basic decision tree classifier does not require any domain knowledge or parameter setting, they can be used for exploratory analysis of the data.

However, in order to build the most accurate model used for prediction making some manipulation is always required (Han, Pei and Kamber, 2011).

Three main algorithms used in Decision Trees are ID3, C4.5, and CART. ID3 (Iterative Dichotomiser 3) and C4.5 were developed and described by Quinlan (1986 and 1993), and expanded in earlier work by Hunt *et.al.* (1966). Both algorithms use Entropy or Information Gain to decide on the attribute selection for the split. Han,

Pei and Kamber (2011) identify C4.5 as the benchmark algorithm to which other supervised learning algorithms should be compared.



**Figure 2.4 Potential ID3-generated Decision Tree<sup>2</sup>**

The last algorithm - CART (Classification and Regression Trees) although developed independently in 1984, follows similar approach by using Gini Impurity and Information Gain for the process of learning decision trees (Breiman, 2017).

Additionally, all algorithms: ID3, C4.5, and CART adopt a greedy approach, where the trees are constructed in a top-down manner. However, while the ID3 and C4.5 allow for multi-way splits (where two or more branches are grown from a node), CART Gini Index selection measure, enforces the binary tree production (Han, Pei and Kamber, 2011).

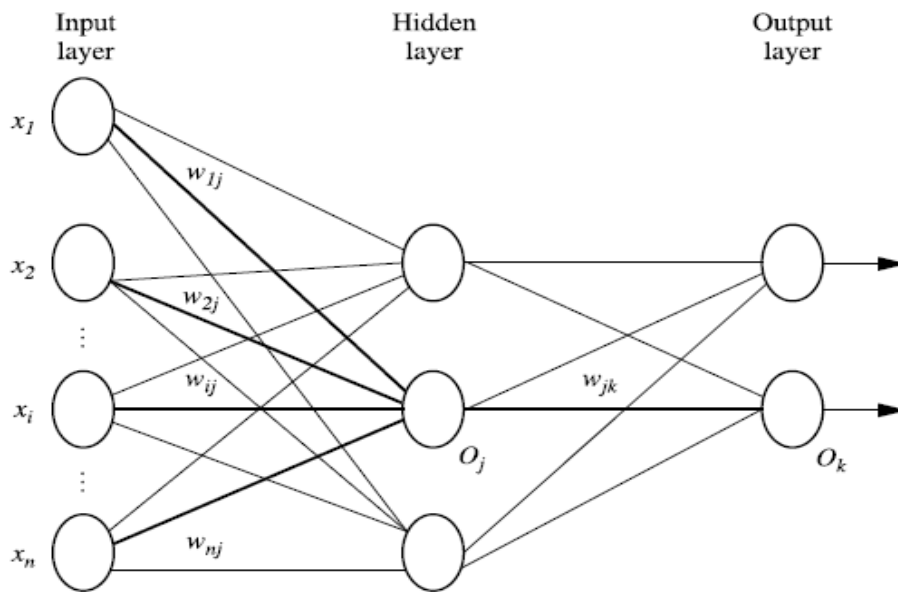
---

<sup>2</sup> Wikivisually.com (2017) Incremental decision tree. Retrieved from: [https://wikivisually.com/wiki/Incremental\\_decision\\_tree](https://wikivisually.com/wiki/Incremental_decision_tree)

### 2.2.3. Neural Networks

Neural networks were originally studied by psychologists and neurobiologists, including previously discussed work by Rosenblatt (1961).

The model of neural network can be described as a set of connected input and output units, where each connection has an associated weight. In each Neural Network one Input and one Output layer exist, while the amount of hidden layers can vary. The outputs of any hidden layer can be inputs to another hidden layer or the output layer. During the learning phase, the weights are adjusted using back-propagation algorithm until the best accuracy of prediction is achieved in production (Han, Pei and Kamber, 2011). The multilayer neural network presented in Figure 2.3 is an example of two-layered network. Only the output and hidden layer units are included in the count, while the input layer is excluded.



**Figure 2.5 Two-layer Neural Network (Han, Pei and Kamber, 2011)**

### 2.2.4. Multiclass Classification Problem

Multiclass classification problem is one which involves classification of instances into one of multiple possible target classes. As the multiclass learning problem is very often related to a real-life scenarios, various approaches were developed, which allow for the

classification. The most common approach is reduction to binary, which includes one-versus-one and one-versus-all approaches (Han, Pei and Kamber, 2011), however extension from binary is a new approach which can be taken (Aly, 2005).

One-versus-one approach involves training of  $K(K - 1) / 2$  binary classifier, where  $K$  is the amount of classes. Each binary classifier is built using a pair of classes from the original data, and then final prediction is made using combined output from multiple binary classifiers. The one-versus-all approach involves creation and training of models per binary class, where one original class is reduced as positive, while all the other as negatives. (Rocha and Goldenstein, 2014). Both these reduction techniques were also analysed in detail by Dietterich and Bakiri (1995), as well as Allwein, Schapire and Singer (2000), while their limitations were presented in the research by Daniely et al. (2011) and Daniely, Sabato & Shwartz (2012).

As mentioned earlier, it is possible to make predictions for the multiclass classification problem by extending some algorithms from the binary classification to multiclass. Research by Aly (2005) provides an evidence of successful implementation of this approach for both Decision Trees and Neural Networks. The research argues that both algorithms can naturally handle both binary or multiclass classification problems. In each case Decision Trees the leaf nodes can simply refer to any of the  $K$  classes to be predicted. At the same time, MultiLayer Neural Networks evolve from having just one neuron in the output layer, with binary output, to having  $K$  binary neurons (Aly, 2005).

As both algorithms of interest of this research, Decision Trees and Neural Networks, can be applied using extensions from binary approach for multiclass prediction, this strategy will be selected for the experiment purposes.

### **2.3. Machine Learning in SWB Research**

Some attempts were made to apply Machine Learning models on the subjective well-being related data. Conry, *et.al.* (2011) presents the results of exploration of self-rated health and quality of life data. The research is also concerned with the Irish population, however the data used was obtained from SLÁN 2007 data (national Survey of Lifestyle, Attitudes and Nutrition). The authors used clustering techniques to explore

associations between the data and reported better self – rated quality of life for the respondents with the healthiest habits (i.e. non-smokers, which exercise more). The paper presented a successful implementation of Unsupervised Machine Learning and the associations present between different clusters and mental health.

However, more significant findings are reported by Jaques, et.al. (2016) where the Multi-Task-Learning (MTL) models are compared (while predicting next-day health, stress, and happiness level). Three models were compared, including Multi-Task Multi-Kernel Learning, Hierarchical Bayes with Dirichlet Process Priors and Neural Networks. This research presents that increasing number of layers improves overall performance of Neural Networks, and allows the model to outperform the other two listed, which supports the selection of the model for the experiment conducted as a part of this research.

#### 2.3.1. Gaps in Research

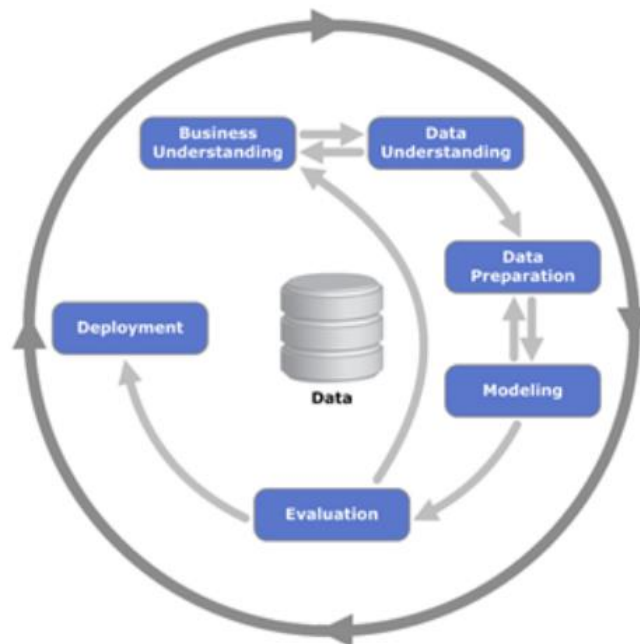
Review of existing literature provides significant evidence and analysis of correlations and causality between individual independent variables (or even selective groups of individual independent variables) and the target variable, which resulted from the vast research performed by economists, psychologist and sociologists. However, the amount of research related to application of Machine Learning in the concept research is scarce.

The research done by Jaques, *et.al.*(2016) is the only one considering a comparison of accuracy of prediction of target variable using Machine Learning classification techniques. However, this research focus is on Multi-Task-Learning models. Therefore, it would be of interest to compare two models: Decision Tree, using C4.5, which was identified as the benchmark algorithm to which other supervised learning algorithms should be compared (Han, Pei and Kamber, 2011), and a Neural Networks, which Jaques *et.al.* (2016) found to have the highest predictive power. Thus, this research will determine effectiveness and performance of these two models in prediction of SWB.

### 3. DESIGN AND METHODOLOGY

A significant amount of the research related to SWB, and its determinants, is based upon survey answers, where one or more questions ask respondents about their life satisfaction and/or happiness. Additionally, supporting questions are asked regarding income, age, employment, marital status, etc., so that the correlation and, ideally, causality of the various components on SWB can be determined. Statistical methods and tests are usually employed to determine those relationships (Benjamin, *et. al.*, 2014; Gerdtham and Johannesson, 2001; Stoica, 2015; Dolan & Metcalfe, 2012).

This chapter describes the general strategy in which research experiment, also based upon survey answers, however involving application of Machine Learning algorithms, will be undertaken. A set of steps and methods to be used is identified here, including data pre-processing and modelling steps, as well as description of evaluation methods and multiple software applications used for the purpose of conducting different stages of the experiment.



**Figure 3.1 Phases of the CRISP-DM reference model (Chapman, *et. al.* 2000)**

Individual steps of experiment plan were created by adapting The Cross Industry Standard Process for Data Mining (CRISP - DM) (Chapman, *et. al.* 2000). The standard approach of CRISP-DM, presented in Figure 3.1, was modified to meet the

need of both this paper and the experiment itself. As the idea behind the approach is that the sequence of the individual phases is not fixed and that going back and forth between them is not only advised, but necessary. The experiment was designed in such a way to benefit the most from this guideline. Figure 3.2 presents the steps of the experiment design, including its sub-tasks. The main phases of Business Understanding, Data Understanding, Data Preparation, Modelling and Evaluation are shared between the two figures; however deployment was removed from the modified approach chosen for this experiment. The approach taken allows for the outcome of each phase sub-task to determine which phase, or phase sub-task, should be performed next. Individual sections of this chapter correspond to the phases, and their subtasks, listed in Figure 3.2 and will provide details regarding each phase of the experiment.

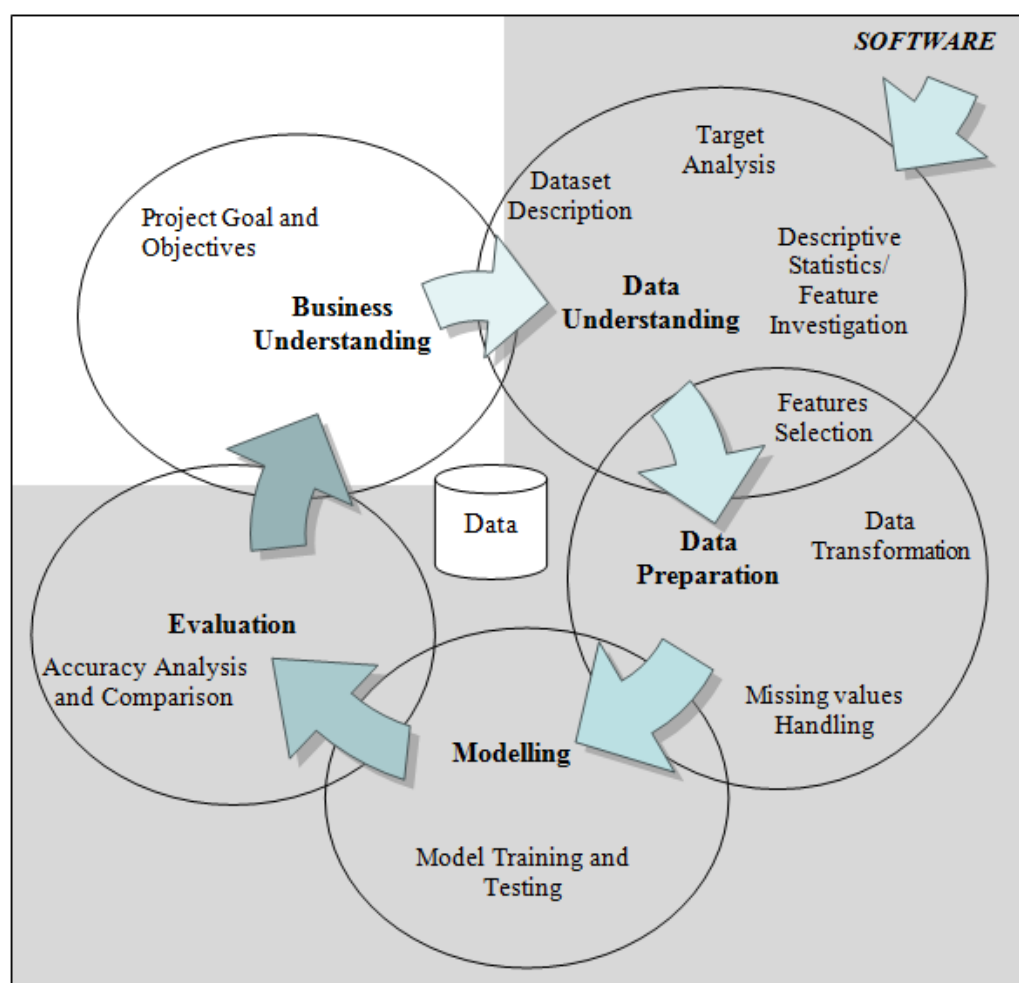


Figure 3.2 High-level Experiment Design Phases (by author)



### **3.1. Business Understanding**

The purpose of the research is to build predictive models based on historical data collected as a part of Healthy Ireland Survey and compare them. The main goal is to test the hypothesis that, while different classifying algorithms can be used for the prediction of subjective well-being, a statistically significant difference exists in the value of prediction accuracy between the models. Thus, an experiment must be conducted. In order to layout methodology accurately it is crucial to begin with highlighting the research question, which then allow for listing research objectives associated with it.

The main research question of this project can be defined as: *Which of the classifiers: Decision Trees or Neural Networks, is more accurate in predicting subjective 'well-being' with the use of specified economic, social and health related factors?*

Thus, following hypothesis can be considered to allow for addressing above research question: *H0: There is a statistically significant difference in the value of prediction accuracy of the subjective well –being between Neural Networks and Decision Trees with  $p$ -value  $< 0.01$ .*

### **3.2. Data Understanding**

This section presents the data understanding phase steps, which enable to determine the quality of the data, and select appropriate strategies to be implemented in the data preparation phase.

#### **3.2.1. Dataset Description**

The data used for the research was obtained from The Irish Social Science Data Archive (ISSDA) and contains 7539 responses to the Healthy Ireland Survey 2015<sup>3</sup>, which was approved by the Research Ethics Committee at the Royal College of Physicians of Ireland (Department of Health, 2016). The survey was carried out

---

<sup>3</sup> Accessed via the Irish Social Science Data Archive - [www.ucd.ie/issda](http://www.ucd.ie/issda)

between late 2015 and early 2016, while a report of its findings was published in October 2016 (Ipsos, 2016).

A representative sample of the Irish population aged 15 and over was achieved by implementation of a multi-stage probability sampling process, in which interviewers were asked to visit pre-selected addresses and then interview a randomly selected individual living under it. The use of this approach allowed every member of the defined population to have the same calculable chance of being included (Department of Health, 2016).

A full survey included questions regarding participants' demographics, education, employment and marital status, but focused mainly on areas such as: general and mental health, and lifestyle factors such as smoking and alcohol consumption, physical activity, nutrition and diet. Additionally, participants interviewed for the survey were asked to complete physical measurement module, in which one's height, weight and waist circumference were recorded. This module was completed by 6,142 respondents (81% of overall study population) (Ipsos, 2016).

### 3.2.2. Target Variable Investigation

The target variable selected for the experiment is a multiclass variable, which involves presence of multiple possible target classes Table 3.1 below presents the exact question and answers from the survey. The values below are stored in the dataset in a coded format represented by numeric values in the range from 1 to 6.

Q45H	Q.45 How much time during the past 4 weeks...Have you been a happy person?	1	All of the time
		2	Most of the time
		3	A good bit of the time
		4	Some of the time
		5	A little of the time
		6	None of the time

**Table 3.1 Target variable question and answers details**

Target variable investigation to be performed should include statistical analysis, with the main focus on the distribution of the individual classes within the target. As the

variable is a multiclass one, it is possible that an imbalance may exist between individual classes, which would have to be addressed in the data preparation phase.

### 3.2.3. Feature Investigation

All independent variables are continuous and stored in the numeric format. Thus, descriptive statistics, such as mean, skewness and kurtosis, will be produced and analysis will be performed in order to verify data distribution, and in effect see if the data normalisation and/or standardisation is required.

Additionally, as the data is secondary and was already used for statistical purposes, no format issues or duplicate values are expected. However, as the data comes from survey, it is possible that some of the participants refused to answer specific questions. Thus, missing values analysis should be performed in order to identify any variables with this issue and select appropriate strategy for addressing it.

### 3.2.4. Feature selection

The selection of attributes is critically important for successful and meaningful modelling of the problem. Guyon and Elisseeff (2003), and Karegowda *et.al.* (2010) argue that inclusion of the redundant attributes not only may be misleading to the algorithms, but can also result in model over-fitting, which in effect may reduce the predictive power of the models built and cripple their overall accuracy. Thus, in the data preparation phase any redundant and/or irrelevant attributes from the dataset will be removed. The selection will be made using output from two feature selection methods: Information Gain ratio and Correlation based feature selection (Frank *et.al.* 2009). Both methods are provided in WEKA Explorer and use the “ranker” search method, which sorts features according to their evaluation. (Karegowda *et.al.* 2010). The first one relies on calculation of the information gain (or entropy) for each feature for the output variable. The values here are always in the range from 0 to 1, where 0 means no information, and 1 means maximum information. The attributes with the highest information gain value are then selected. The second method uses Pearson’s correlation coefficient. After the correlation between each attribute and the target variable is calculated, only those attributes with the highest positive or negative

correlation are kept and those with a lowest correlation (value close to 0) are dropped. (Frank *et.al.*, 2009)

### **3.3. Data preparation**

This section describes any tasks related to the data preparation phase, which includes missing values handling, feature selection, and data transformation like imbalance removal and normalisation.

#### **3.3.1. Missing Values Handling**

Any missing values identified in the feature investigation step will be addressed, as it was proved that the existence of missing values affects the performance of some Machine Learning classifiers (Pelckmans, Brabanter., Suykens, and Moor, 2005). The strategy for replacing missing values will be determined based on the results percentage missing. All variables with the count of missing >40% of overall count of observations will be removed. This threshold was selected based on the research done by Silipo, Adae and Hart (2015), where it proved to lead to the best accuracy achieved by the Machine Learning models. For the remaining variables, if they are selected as features for the model building phase, MCAR test will be performed, and the decision about an imputation of missing values will be made based on it.

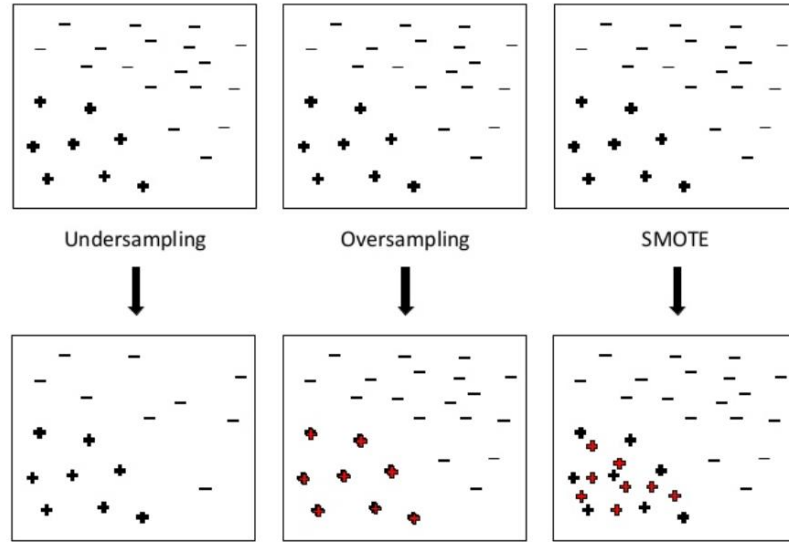
#### **3.3.2. Feature Selection**

As previously discussed in the Data Preparation section of this chapter, two techniques provided by WEKA will be used in order to perform Feature Selection: Correlation Based Feature Selection and Information Gain Based Feature Selection (Frank *et.al.*, 2009). The variable list produced by both will be compared and all variables present on either of the list will be included in the final clean dataset.

#### **3.3.3. Imbalance reduction**

As previously discussed, the target is a multiclass variable, thus it is possible that an imbalance may exist between individual classes. Papers by Weiss and Provost (2001), and Chawla, Japkowicz, and Kotcz (2004), and He and Garcia (2009) document that

for multiple classifiers, imbalance removal leads to overall improvement in classification performance. The sampling methods, which may be used for resolving imbalanced dataset issue, include two main random sampling techniques: over-sampling and under-sampling, as well as SMOTE. All of those techniques are presented in Figure 3.3 (Pozzolo, 2016).



**Figure 3.3 Imbalance removal techniques (Pozzolo, 2016)**

Batista, *et.al.* (2005) defined Random Undersampling as a method, which aims to balance out target class distribution by random elimination of observations from the majority class. Respectively, Random Oversampling was defined as a method, which aims to balance out distribution of target class via random replication of instances in the minority class examples. Drawbacks of both techniques were also described. For Random Undersampling, it is that the use of this method can lead to deletion of potentially useful data, which could be significant for training of a model. For Random Oversampling, it is that the method can increase the likelihood of model over-fitting, as all the new instances created are always exact copies of the existing observations in the minority class. Therefore, the predictions made by the model constructed, are not really accurate, as they are made for the same one replicated instance. Additionally, all the previously mentioned papers (Weiss and Provost (2001), Chawla, Japkowicz, and Kotcz (2004), He and Garcia (2009) and Batista, *et.al.* (2005)) claim SMOTE (Synthetic Minority Over-Sampling) Technique offers an alternative to the two previously discussed. SMOTE is also an over-sampling method; however the main

idea behind it is to create new minority class instances by interpolating between several examples from original minority class data. Thus, the method doesn't risk information loss, or over-fitting of the models, and will be the one used for imbalance reduction, if required.

#### 3.3.4. Normalisation

The final task in the data pre-processing phase will be normalisation, which was proven to improve the accuracy and efficiency of Machine Learning algorithms such as Neural Networks, K-nn and rule based learners (Shalabi and Shaaban, 2006). Most common data normalisation methods include:

- **Min-max normalisation** - a technique, which normalises the data through application of a linear transformation and scaling it to the range of 0 to 1. The computation formula for Min-max normalisation is defined as:

$$v' = (v - \text{min}) / (\text{max} - \text{min}) * (\text{newmax} - \text{newmin}) + \text{newmin}$$

where:

$v$  = old variable

$v'$  = transformed variable.

newmin = minimum of the normalised dataset

newmax = maximum of the normalised dataset

- **Z-score normalisation** - is a technique, in which the values are normalised based on the mean and standard deviation of an attribute. Thus, normalisation formula for value  $v$  into  $v'$  is:

$$v' = ((v - l) / \text{std})$$

where:

$l$  = mean

$\text{std}$  = standard

- **Decimal scaling** – this technique normalises the data by moving the decimal point of values, which depends on the maximum absolute value of an attribute. A normalised value  $v'$  is therefore produced by computing:

$$v' = (v / 10^j)$$

where:

$j = \text{smallest integer such that } \text{Max}(|v'|) < 1.$

The experimental results (Shalabi and Shaaban, 2006) suggest choosing the min-max normalisation method, as it proved to have the highest positive effect on the performance of all the machine learning algorithms tested, including Neural Networks and Decision Trees, which are in scope of this research study.

### **3.4. Modelling**

The purpose of this research is to investigate and analyse in detail application of two Supervised Machine Learning techniques: Decision Trees and Neural Networks, and to compare the accuracy of predictions made by each of these models.

As previously discussed, multiple Supervised Machine Learning techniques exist, however the amount of the previous research on application of Supervised Machine Learning is scarce. Therefore, Decision Trees were selected for the purposes of this research, as they are the most fundamental machine learning models, which are able to provide interpretability and information about the importance of individual features. Additionally, Han, Pei and Kamber (2011) identify C4.5 used by the Decision Trees as the benchmark algorithm to which all other supervised machine learning algorithms should be compared. At the same time, Neural Networks were proven to have high predictive power (Jaques *et.al.* 2016), especially with multiclass categorisation (Aly, 2005).

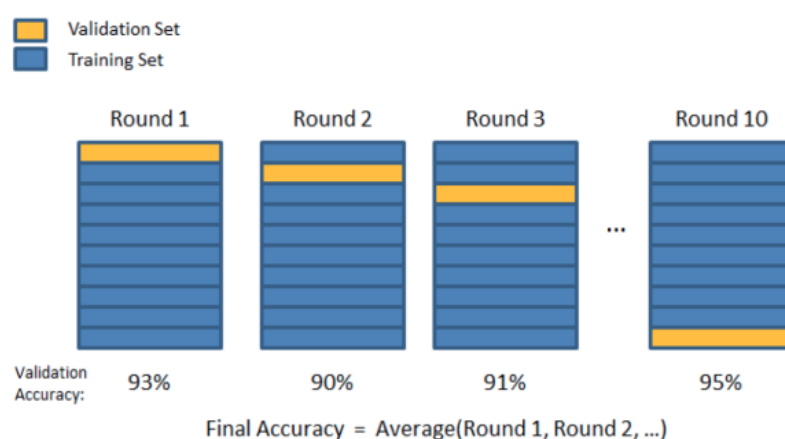
Before moving into model comparison and evaluation phase each of the above models will be tuned and adjusted for the best performance. This step will be performed using full dataset split into 70% for training and 30% for validation parts. The best model will be chosen based on The Misclassification Rate, Average Squared Error and ROC index values produced.

### **3.5. Evaluation**

This section of the reports presents the steps planned for the evaluation phase of the experiment, which include: the comparison of the model performance and the testing for statistical significance leading to hypothesis acceptance and/or rejection.

### 3.5.1. Model performance comparison

Model evaluation will be performed using the results gathered during the k-fold cross validation step, with  $k = 10$  (Refaeilzadeh, Tang and Liu, 2009). As the purpose of this research is to evaluate and compare the accuracy of two predictive models, stratified k-fold cross validation will be used for the individual models validation (Moreno-Torres, Sáez and Herrera, 2012). K-fold cross validation allows to test the predictive accuracy of the model using training data only and without biasing the prediction (Bengio and Grandvalet, 2004). It achieves this through division of data into K equal subsets followed by iterative creation and testing of predictive models. Each time one of the subsets is withheld and used for the testing of the model, while the remaining folds are used for training (Refaeilzadeh, Tang and Liu, 2009). This means that the use of 10-fold cross validations automatically enforces 90%/10% split in Training and Test Sets respectively. The average results from the k-folds will be then taken to produce single overall result. Additionally, a stratified version of this method was selected, as it maintains the proportion of classes present in the target at the whole population level in all the individual folds created (Moreno-Torres, Sáez and Herrera, 2012). Figure 3.4 presents schematic representation of 10-fold cross validation.



**Figure 3.4 Schematic representation of 10-fold cross validation**

The fit statistics outputs produced in the cross-validation process include numerous metrics. The choice of the fit statistic to be used depends from the prediction of interest. In overall all the metrics can be grouped as follows:



Prediction Type	Fit Statistic	Direction
Decisions	Misclassification	Smallest
	Average Profit/Loss	largest/smallest
	Kolmogorov-Smirnov Statistic	Largest
Rankings	ROC Index (concordance)	Largest
	Gini Coefficient	Largest
Estimates	Average Squared Error	Smallest
	Schwarz's Bayesian Criterion	Smallest
	Log-Likelihood	Largest

**Table 3.2 Fit Statistics Grouping by Prediction Type**

Therefore, for the purposes of this research, where the prediction type falls into the ‘decision’ category, the results generated will be analysed using the misclassification rate value produced for each model.

The most common metric used for the model effectiveness assessment is the accuracy, which is computed as (Costa *et.al.*, 2007):

$$Accuracy = (TP + TN)/(TP + FP + FN + TN)$$

Where:

TP = True Positive (number of positive instances classified as positive)

FP = False Positive (number of negative instances classified as positive)

FN = False Negative (number of positive instances classified as negative)

TN = True Negative (number of negative instances classified as negative)

However, it is also possible to evaluate model performance using Misclassification Rate instead, also known as Error Rate (Costa *et.al.*, 2007). While Accuracy shows how often the classifier is correct, Misclassification Rate presents the figure on how often the classifier is wrong. Thus, there are 2 ways of performing its computation, first one being (Costa *et.al.*, 2007):

$$Misclassification\ Rate = (FP+FN)/(TP + FP + FN + TN)$$

Where:

TP = True Positive Rate (number of positive instances classified as positive)

FP = False Positive Rate (number of negative instances classified as positive)  
 FN = False Negative Rate (number of positive instances classified as negative)  
 TN = True Negative Rate (number of negative instances classified as negative)

The second option is to simply calculate Misclassification Rate as (Costa *et.al.* 2007):

$$\text{Misclassification Rate} = 1 - \text{Accuracy}.$$

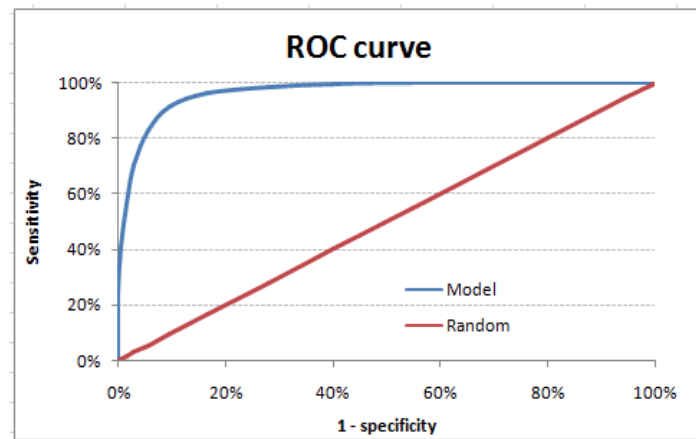
Both metrics use Confusion Matrix (Figure 3.6) as a primary source of data.

		Predicted	
		Condition positive	Condition negative
Actual	Condition positive	True positive (TP)	False negative (FN)
	Condition negative	False Positive (FP)	True negative (TN)

**Figure 3.5 Confusion Matrix**

Misclassification Rate comparison will be the primary determinant of model performance, however, in the model adjustment step, it will be additionally supported by the analysis of two other Fit Statistic: Averaged Squared Error and ROC index (which reflects AUC - the area under the ROC curve). Those two will not be used for final model comparison or tested for the purpose of accepting or rejecting the hypothesis, but will provide additional insight when building and tuning the models. For example, the AUC value can be generalized into following model performance groups (Bradley, 1997):

- .90-1 = excellent
- .80-.90 = good
- .70-.80 = fair
- .60-.70 = poor
- .50-.60 = fail



**Figure 3.6 Example ROC curve (Bradley, 1997)**

### 3.5.2. Statistical significance and hypothesis evaluation

Finally, a statistical significance of difference in model performance for each of the experiment results will be verified in order to accept or reject the hypotheses stated. As the distribution of the experiment results may not be normal, the test used to verify a statistical significance (with p-value set to 0.01) will be Wilcoxon Signed-Rank Test (Gibbons and Chakraborti, 2011).

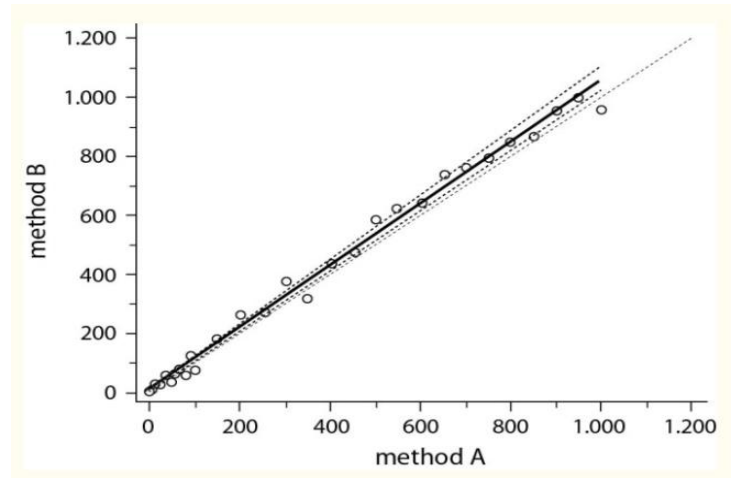
The Wilcoxon Signed-Rank Test is a non-parametric statistical hypothesis test, which should be used if a normal distribution of the population tested can't be assumed. The test is a paired difference test based on rank and can be used to compare:

- two related samples,
- matched samples,
- repeated measurements on a single sample to assess (Rey and Neuhaus, 2011).

The Wilcoxon Signed-Rank Test will be additionally supported by the Agreement Plot, also called Bland-Altman plot (Giavarina, 2015), in which a Decision Tree results for Misclassification Rate will be plotted against a Neural Networks Misclassification Rates.

The Agreement Plot will present a regression line with a slope of 1, which identifies the points where the difference between the values is equal to 0. Thus, it allows for

visualisation of the result points, including the mean, and their position in relation to the regression line, which corresponds to the results of t-test (Giavarina, 2015).



**Figure 3.7 Example Agreement Plot (Giavarina, 2015)**

### **3.6. Software**

The last element to be discussed, as a part of the Design and Methodology section, is a selection of tools used to perform different steps of the experiment. Choosing the right software is important for the successful research project. As the project involves multiple phases, starting with the data understanding and descriptive statistic, through the data pre-processing to the predictive modelling, it is often necessary to perform some tasks or steps using different tools. In the case of this paper following tools were used: WEKA (Hall *et.al.* 2009), SAS Studio and SAS Enterprise Miner (Hall *et.al.*, 2014)

The original dataset was received in *.sas7bdat format*, and requires conversion into *.csv* in order to be used by any other analytical tool than SAS. However, as SAS is a powerful programming language, which allows for efficient manipulation of data, it is planned to use it to perform most of the data exploration and pre-processing tasks. Additionally, the platform itself supports variety of statistical methods through the set of pre-build libraries and functions. This allows for the generation of the descriptive statistics tables and the supporting graphs.

Nevertheless, few steps related to the data preparation will be performed via WEKA. Those will include feature selection and normalisation, as the WEKA software offers much better and straightforward way of performing those actions. However, this will

require implementation of import and export functions within SAS code to allow for the conversion between different file formats at different stages of the data preparation phase.

The modelling phase of the experiment will be performed using SAS Enterprise Miner, which allows for building predictive and descriptive models, and their comparative analysis. The tool supports multiple algorithms and techniques, including Decision Trees, Naïve Bayes, Regression, SVM and Neural Networks (Hall *et.al*, 2014).

### **3.7. *Strength and limitations***

Experiment design and methodology presented highlight not only the actions that must be performed as a part of each phase, but also many possible issues, which may occur, together with the solutions to be implemented in order to achieve robust modeling results.

Firstly, two different types of models will be trained and tested in order to obtain insight on usefulness of the survey data in prediction of subjective feeling of happiness. Decision trees are the most fundamental machine learning models, which are able to provide interpretability and the information about the importance of the individual features. Neural Networks were previously proven to have high predictive power (Jaques *et.al*. 2016), and performing well with multiclass target categorisation (Aly, 2005).

As the models will be trained and tested using the stratified 10-fold cross validation, the predictions obtained should not only be representative, but also more accurate. Additionally, the results obtained from 10 iterations will be sufficient for comparison of models performance and hypothesis testing.

The major limitation for the research is the presence of imbalance in the multiclass target and fairly low correlation between the feature and the target variables, which may result in the low prediction accuracy of both models tested. Additionally, the use of Synthetic Minority Over-sampling Technique on minority class may produce instances different than the real world data, which in effect may skew model results.

## **4. IMPLEMENTATION AND RESULTS**

This chapter presents the practical implementation of the experiment design discussed in Chapter 3 of this paper. However, for the practical purposes the Data Understanding and Data Preparation sections from previous chapter were merged. This allows for more accurate presentation on how individual data understanding steps were directly followed by specific data preparation steps.

### ***4.1. Business understanding***

The purpose of the research, as previously discussed, is to build two predictive models using historical data collected as a part of Healthy Ireland Survey, and provide a proof that, while different classifying algorithms can be used for the prediction of subjective well-being, a statistically significant difference exists in the value of prediction accuracy between the models.

An experiment conducted to achieve above goal included following phases:

- 1) Exploratory analysis of the Healthy Ireland Survey dataset, including the target variable and independent variables
- 2) Data preparation, including missing values handling, feature selection, data resampling and normalisation
- 3) Modelling, including models selection, training and testing
- 4) Evaluation of results

### ***4.2. Data Understanding and Preparation***

Healthy Ireland Survey dataset is a flat file, which consist 169 variables and 7,539 instances. The purpose of the initial data quality investigation is to identify any potential issues by the analysis of descriptive statistics, trends in data, its distribution, missing values and/or outliers. This includes both the dependant variable, and the independent variables.

#### 4.2.1. Target Variable

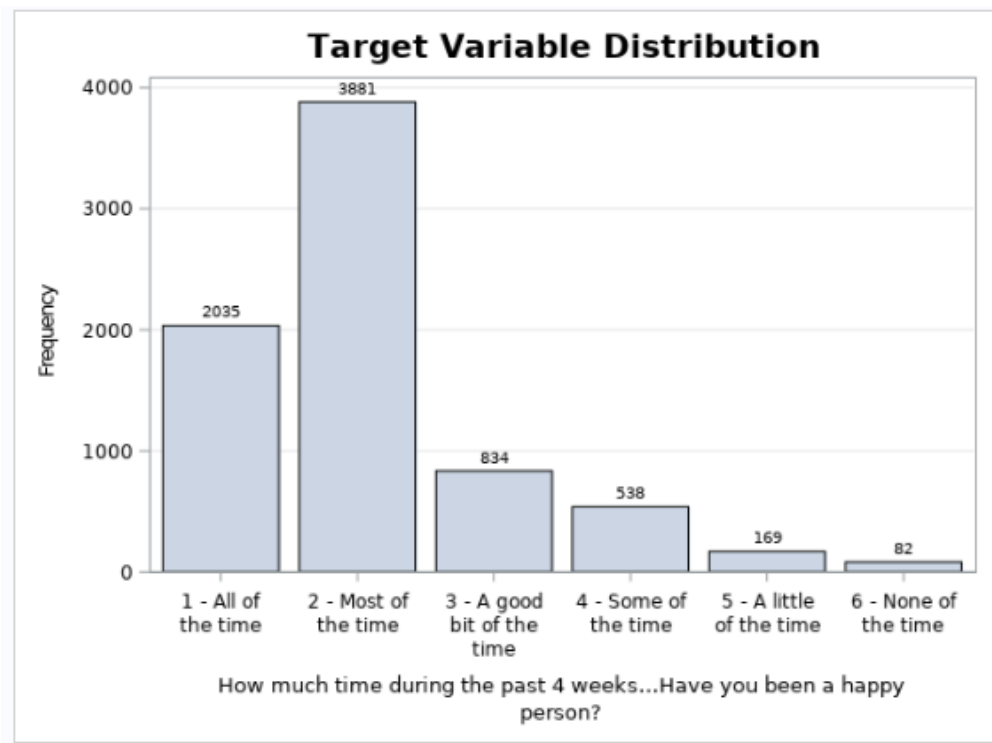
For the purposes of this research one of the variables in the dataset was selected as the target. The variable selected captures respondents answer regarding the subjective feeling of happiness, or rather the amount of the time of feeling ‘happy’, which respondent reports. The variable had no missing values and contained answer to the question regarding the subjective feeling of happiness. Table 3.1 in the Design and Methodology section of this document presents the detailed information regarding the question and possible answers to be given. One of the goals of this research is to correctly predict the answer based on the other survey data. As six possible answers are available, the prediction of the target variable value is described as multiclass classification problem. This will have to be taken into consideration when moving into modelling phase and making a selection of the algorithm to be used.

Answers to the question in Table 3.1 are captured in the dataset in the numeric format with the values from 1 to 6. For the purposes of the exploratory analysis the coded values were mapped as per Table 4.1.:

Code	Class
1	1 - All of the time
2	2 - Most of the time
3	3 - A good bit of the time
4	4 - Some of the time
5	5 - A little of the time
6	6 - None of the time

**Table 4.1 Target variable mapping**

The mapping performed included adding a code number to the original answer, which allowed for better visualisation and analysis, which has shown that the distribution of observations in the target variable, presented in the Figure 4.1 below, presents significant imbalance of the distribution of individual classes within the target variable, which would have to be addressed before moving into modelling phase of the experiment.



**Figure 4.1 Target variable classes' distribution**

#### 4.2.2. Independent variables – missing values handling

As mentioned at the beginning of this section, the total variable count for the dataset was equal to 169. However, initial analysis of the missing values has shown that 30 variables exist, where the missing observations count is greater than 40% of the over observation count. Considering the research by Silipo, Aday and Hart (2015) and an unequal distribution of the target variable, a decision was made to remove those variables from the dataset in order to prevent them influencing the statistics and models themselves. Table 4.2 presents full list of the removed variables, together with the associated label and the value of percentage of observations missing at the time of removal. Most of the variables removed were identified as the supporting variables for the main question. For example, if respondent is a 'non-smoker' all the questions related to smoking frequency will be left blank.



<i>Var</i>	<i>Label</i>	<i>% missing</i>
<b>iq5d</b>	How often in the last 4 weeks did you consult nurse working within a GP practice on your own behalf excluding visits where you also consulted the GP?	66.44
<b>iq5f</b>	How many times have you consulted medical consultant in the past 4 weeks?	70.66
<b>iq9a1</b>	On average how many of the Manufactured cigarettes do you smoke each day	81.83
<b>iq9a2</b>	On average how many of the Hand-rolled cigarettes do you smoke each day	81.83
<b>iq9a3</b>	On average how many of the Pipes full of tobacco do you smoke each day	81.83
<b>iq9a4</b>	On average how many of the Cigars do you smoke each day	81.83
<b>iq9a5</b>	On average how many of the Other tobacco products do you smoke each day	81.83
<b>iq9b1</b>	On average how many of the Manufactured cigarettes do you smoke each week	96.17
<b>iq9b2</b>	On average how many of the following tobacco products do you smoke each week Hand-rolled cigarettes	96.17
<b>iq9b3</b>	On average how many of the pipes full of tobacco products do you smoke each week	96.17
<b>iq9b4</b>	On average how many of the Cigars do you smoke each week	96.17
<b>iq9b5</b>	On average how many of the Others tobacco products do you smoke each week	96.17
<b>niq32</b>	How much time did you spend doing vigorous physical activities on one of those days?	67.04
<b>q11</b>	During the past 12 months have you stopped smoking for one day or longer because you were trying to quit smoking?	74.36
<b>q12_1</b>	During your last attempt to give up did you use any help? - Nicotine patches, gum, lozenges, spray	87.16
<b>q12_10</b>	During your last attempt to give up did you use any help? - Don't Know	87.16
<b>q12_11</b>	During your last attempt to give up did you use any help? - Refused	87.16
<b>q12_2</b>	During your last attempt to give up did you use any help? - Varenicline/Champix or Bupropion/Zyban (prescribed medication)	87.16
<b>q12_3</b>	During your last attempt to give up did you use any help? - Acupuncture	87.16
<b>q12_4</b>	During your last attempt to give up did you use any help? - Smokers telephone Quitline/Helpline	87.16
<b>q12_5</b>	During your last attempt to give up did you use any help? - www.quit.ie	87.16
<b>q12_6</b>	During your last attempt to give up did you use any help? - www.facebook.com/HSEquit	87.16
<b>q12_7</b>	During your last attempt to give up did you use any help? - E-cigarettes	87.16
<b>q12_8</b>	During your last attempt to give up did you use any help? - Other aid, help, support	87.16
<b>q12_9</b>	During your last attempt to give up did you use any help? - No help used	87.16
<b>q13</b>	Are you currently...?	77.99
<b>q58_2</b>	How would the chief income earner define their current situation with regard their work?	68.17
<b>q59b</b>	How many hours per week?	91.33
<b>q8</b>	About how long has it been since you last smoked tobacco products?	70.38
<b>slq9b</b>	Non smoker	99.27

**Table 4.2 Variables rejected due to missing values count**

#### 4.2.3. Independent variables – feature selection

The removal of the variables with the majority of missing instances contributed to the reduction of overall variable count to 139. Additionally, 2 other variables were removed, as their value was a result of the derivation performed using target variable as an input (Ipsos, 2016). Those were:

- PMHP group - Positive mental health measurement
- High EVI group - High Energy and Vitality group based on the PMHP score.

This brought the total number of the variables to 137, which would still have to be decreased in order to achieve the project goal and be able to perform modeling. Thus, dimensionality reduction, in the form of feature selection, was implemented in order to decrease the number of the variables to consider.

Feature selection, which allows the models for much easier and faster data analysis, was performed using two methods available in WEKA:

- Correlation Based Feature Selection
- Information Gain Based Feature Selection (Frank *et.al.* 2009).

As discussed in the Design and Methodology chapter, the first method uses Pearson's correlation coefficient and drops those attributes with the lowest correlation value (closest to 0). The other method uses information gain value (entropy) and drops the variables with the lowest score. Both methods use a ranker search method, where a specific value of threshold must be provided. For the purposes of this experiment and research the threshold value was left with a default value of -1.7976931348623157E308. Both methods were setup to output top 20 values.

The execution of both methods resulted in the selection of slightly different lists of 20 variables. 17 variables selected were present in both outputs, however in the different order, due to different rank given to variables by correlation and/or the information gain method. 6 variables different were: niq37, q5e and q46sp\_16 for Information Gain Based Feature Selection, and q44b, q43 and q44c for Correlation Based Feature Selection. The decision was made to keep all the features selected by both methods (total of 23 features) and use them for the predictive models creation.

Table 4.3 below presents all the features selected and categorises them according to previously discussed groups of factors affecting subjective well-being

<b>Health and Health related habits</b>		
<b>General Health</b>	<b>Mental Health</b>	<b>Diet and Nutrition</b>
spq1 - How is your health in general?	q45a - How much of the time during the past 4 weeks.... Did you feel full of life	q24 - How often do you eat vegetables or salad, excluding juice and potatoes?
q2 - Do you have any long standing illness or health problem i.e. problems which have lasted or will last for at least 6 months or more?	q45b - How much of the time during the past 4 weeks.... Have you been a very nervous person	
	q45d - How much of the time during the past 4 weeks.... Have you felt calm and peaceful	
	q45e - How much of the time during the past 4 weeks.... Did you have a lot of energy	<b>Physical Activity</b>
q3 - For at least the past six months to what extent have you been limited in everyday activities because of health problems i.e. an on-going physical or mental health problem illness or disability?	q45g - How much of the time during the past 4 weeks.... Did you feel worn out	q31 - During the last 7 days on how many days did you do vigorous physical activities like heavy lifting competitive sport or fast cycling?
	q45i - How much of the time during the past 4 weeks.... Did you feel tired	
	q46sp_7 - Which of these changes if any would you like to make that would improve your health and wellbeing? - Reduce the amount of stress in my life	
q5e - When was the last time you consulted a medical or surgical consultant on your own behalf?	q46sp_8 - Which of these changes if any would you like to make that would improve your health and wellbeing? - Sleep better	niq37 - During the last 7 days, how much time did you spend sitting on a weekday?
	q46sp_9 - Which of these changes if any would you like to make that would improve your health and wellbeing? - Relax more	
q54a - Do you have a full medical card?	q46sp_16 - Which of these changes if any would you like to make that would improve your health and wellbeing? - Be more financially secure	sipaq - Standardised Personal Activity Level
<b>Social Relationships</b>		
<b>Social Connectedness</b>		
q43 - Do you participate in any social groups or clubs?		
<b>Economic and Physical safety</b>		
<b>Employment</b>	<b>Vandalism and Crime</b>	
q58 - How would you define your current situation with regard to work?	q44b - How much of a problem are each of the following in your neighbourhood? Graffiti on walls or buildings	
	q44c - How much of a problem are each of the following in your neighbourhood? Vandalism and deliberate damage to property	

**Table 4.3 Selected Features Categorisation**

Table 4.4 present the outputs from individual feature selection methods including the output order and the correlation/information gain metrics associated with each of the selected variables.

Correlation Based Feature Selection			Information Gain Based Feature Selection		
#	Correlation value	Variable name	#	Info. Gain Value	Variable name
1.	0.1822	q45d	1.	0.32768	q45d
2.	0.1674	q45a	2.	0.2307	q45a
3.	0.149	q45e	3.	0.22094	q45e
4.	0.1386	q45b	4.	0.09711	q45i
5.	0.1195	q45g	5.	0.09651	q45g
6.	0.1137	q45i	6.	0.09096	q45b
7.	0.0968	spq1	7.	0.04332	spq1
8.	0.0837	q46sp_7	8.	0.03408	q46sp_7
9.	0.0758	q3	9.	0.0228	q3
10.	0.061	q2	10.	0.01507	q2
11.	0.0505	q46sp_8	11.	0.01271	q46sp_8
12.	0.0497	Sipaq	12.	0.01124	q46sp_9
13.	0.0476	q44c	13.	0.01109	q58
14.	0.0465	q54a	14.	0.01034	sipaq
15.	0.0434	q43	15.	0.00947	q31
16.	0.0434	q46sp_9	16.	0.00939	niq37
17.	0.0412	q31	17.	0.00895	q54a
18.	0.0404	q24	18.	0.00863	q46sp_16
19.	0.04	q44b	19.	0.00855	q24
20.	0.0296	q58	20.	0.0085	q5e

**Table 4.4 Feature selection algorithms output comparison**

#### 4.2.4. Feature investigation

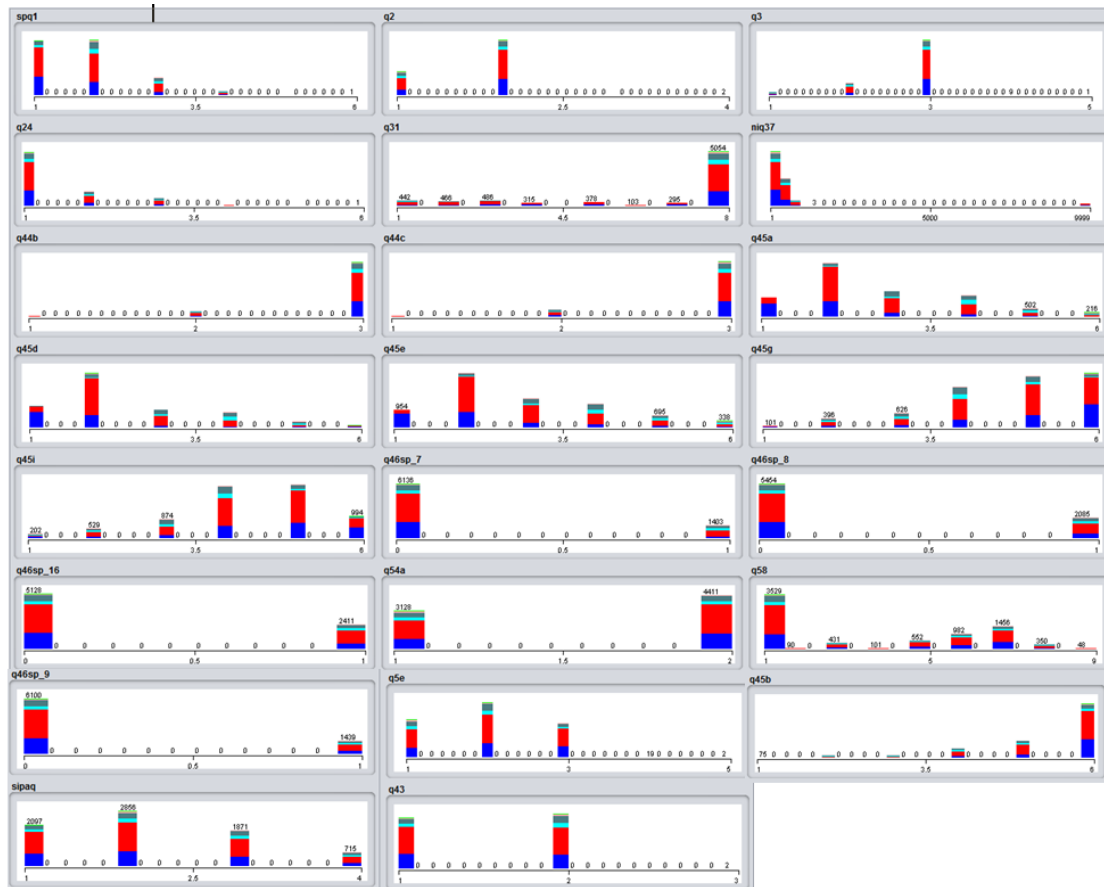
All of the variables selected by both feature selection algorithms present a positive correlation value with the target variable with the range from 0.0296 to 0.1822. This is present due to the nature of the data, and the fact that all values are coded representation of the answers given. The features originated from the following survey question sections: General Health, Mental Health, Diet and Nutrition, Physical Activity and Social Connectedness, and were grouped by using groups indicated in the government publications. Those are: “Health and Health related habits”, “Social Relationships”, and “Economic and Physical Safety” While Table 4.4 presents the

summary statistics, Table 4.3 present the details regarding variables assignment to the categories selected from Eurostat Guidelines (Eurostat, 2015) and subcategories based on Healthy Ireland Survey sections. Eurostat publication was selected as a guideline for categories selection, as it presents the recent statistics on the quality of life of all European Union countries, which includes Ireland. This categorisation table presents that the variables selected correspond to the ones identified in the previous research related to application of the Machine Learning models on the subjective well-being related data (Conry, *et.al.* 2011)

Variable	Mean	Std Dev	Min	Max	Median	Skewness	Kurtosis
q45d	2.45	1.18	1	6	2	0.96	0.43
q45a	2.68	1.26	1	6	2	0.72	-0.13
q45e	2.86	1.34	1	6	3	0.62	-0.42
q45b	5.33	1.08	1	6	6	-1.81	3.02
q45g	4.70	1.22	1	6	5	-0.85	0.19
q45i	4.26	1.20	1	6	4	-0.67	0.16
spq1	1.79	0.81	1	6	2	0.93	0.84
q46sp_7	0.19	0.39	0	1	0	1.61	0.60
q3	2.74	0.53	1	5	3	-1.87	2.75
q2	1.69	0.47	1	4	2	-0.74	-1.12
q46sp_8	0.28	0.45	0	1	0	1.00	-1.00
sipaq	2.16	0.94	1	4	2	0.37	-0.79
q44c	2.84	0.43	1	3	3	-2.73	6.99
q54a	1.59	0.49	1	2	2	-0.35	-1.88
q43	1.52	0.50	1	3	2	-0.07	-1.98
q46sp_9	0.19	0.39	0	1	0	1.57	0.48
q31	6.51	2.39	1	8	8	-1.24	-0.09
q24	1.47	0.81	1	6	1	1.86	3.34
q44b	2.87	0.39	1	3	3	-3.01	8.86
q58	3.65	2.73	1	9	3	0.26	-1.66
niq37	641.62	1750.50	1	9999	300	5.10	24.32
q5e	1.98	0.76	1	5	2	0.08	-1.11
q46sp_16	0.32	0.47	0	1	0	0.77	-1.40

**Table 4.5 Summary Statistics for selected features**

The analysis of summary statistics present in Table 4.5, as well as the distribution presented in Figure 4.2 shows that the range of the values present in the raw data varies. Although the variance is not extreme in most of the cases, one variable exists, which has much greater max value than all the rest. Figure 4.2 clearly present significant scale increase for variable niq37 (“During the last 7 days, how much time did you spend sitting on a weekday?”), where the range of values is from 1 to 9999, while the most of the variables present a range from 1 to 6. Presence of such a discrepancy in the range values of different variables can have an impact on the predictive models, in particular Neural Networks. As discussed earlier in the literature review, Neural Networks present much better accuracy when working with the normalised data (Shalabi and Shaaban, 2006). Thus, normalisation of features was performed before moving into modelling phase in order to allow each feature for approximately proportional contribution.



**Figure 4.2 Feature distribution ranges**

#### 4.2.5. Data normalisation

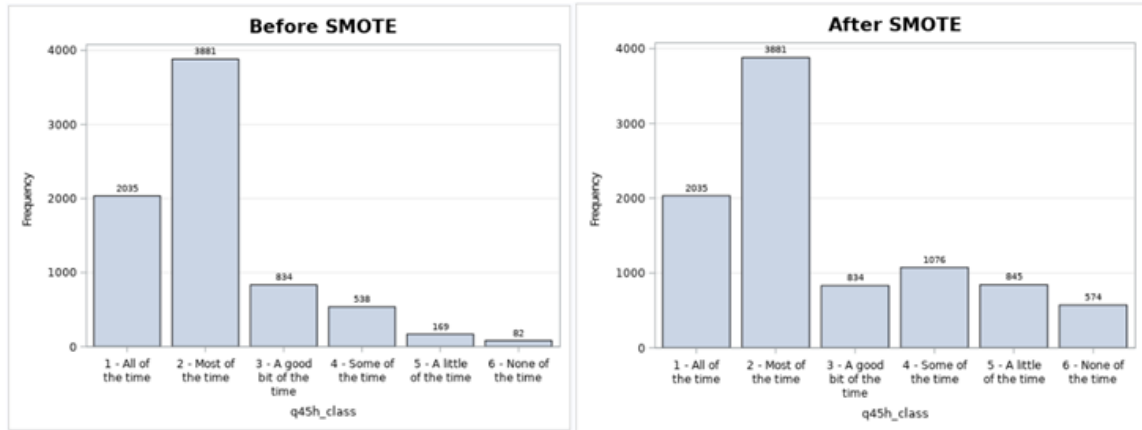
The normalisation was performed in WEKA; in effect all of the selected features were rescaled to fit the range of 0 to 1. Making 1 the largest value for each attribute and 0 the smallest one.

#### 4.2.6. Target variable - imbalance removal

In order to remove the imbalance identified in the distribution of the individual classes in the target variable SMOTE over-sampling was performed. This method was selected, over the other methods previously discussed, as it doesn't lead to information loss and was proven to outperform regular random over-sampling (Batista, *et.al.* 2005). Table 4.6 below presents exact counts of individual classes prior and after application of over-sampling on the dataset.

Label	Observation count in unbalanced dataset	Increase in %	Observation count after SMOTE
1 - All of the time	2035	0	2035
2 - Most of the time	3881	0	3881
3 - A good bit of the time	834	0	834
4 - Some of the time	538	100	1076
5 - A little of the time	169	400	845
6 - None of the time	82	600	574
TOTAL	7539	n/a	9245

**Table 4.6 Count of individual classes before and after over-sampling**



**Figure 4.3 Imbalanced vs. balanced target**

Figure 4.3 above presents the distribution of classes before and after SMOTE application. It is visible here that, while the overall count of the instances for the “2- Most of the time” target class remains unchanged, the count of instances in target classes: “4 - Some of the time”, “5 - A little of the time”, “6 - None of the time” had increased. This change in the counts may have a positive impact on the performance of the models tested, which will be verified by comparing the results of the experiment executed using both the imbalanced and the balanced data

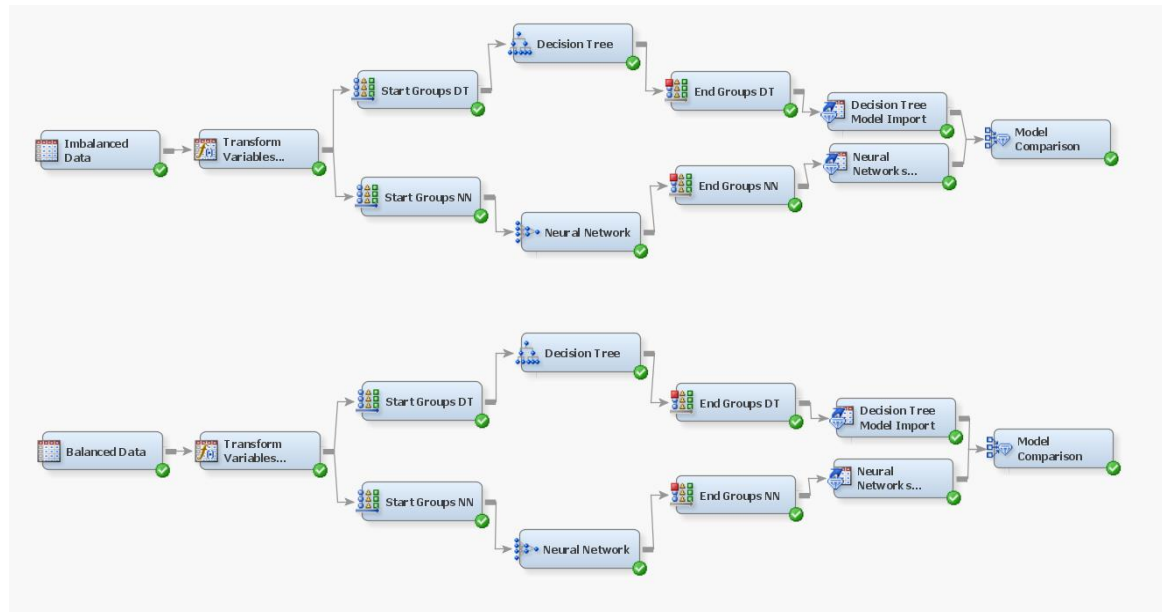
### **4.3. Modelling**

This phase of the research involved the creation and the testing of Decision Tree and Neural Networks classification models to predict the answer regarding subjective feeling of well-being. As previously discussed, all input data was pre-processed and only selected and normalised features are used as inputs for the models build.

Four supervised machine learning models were compared in total and used in 2 separate experiments. The first experiment involved a Decision Tree and Neural Network performance comparison using the dataset, which have undergone all the pre-processing changes except the imbalance removal using SMOTE. The second experiment also involved a Decision Tree and Neural Network performance comparison; however the balanced dataset was used here instead.

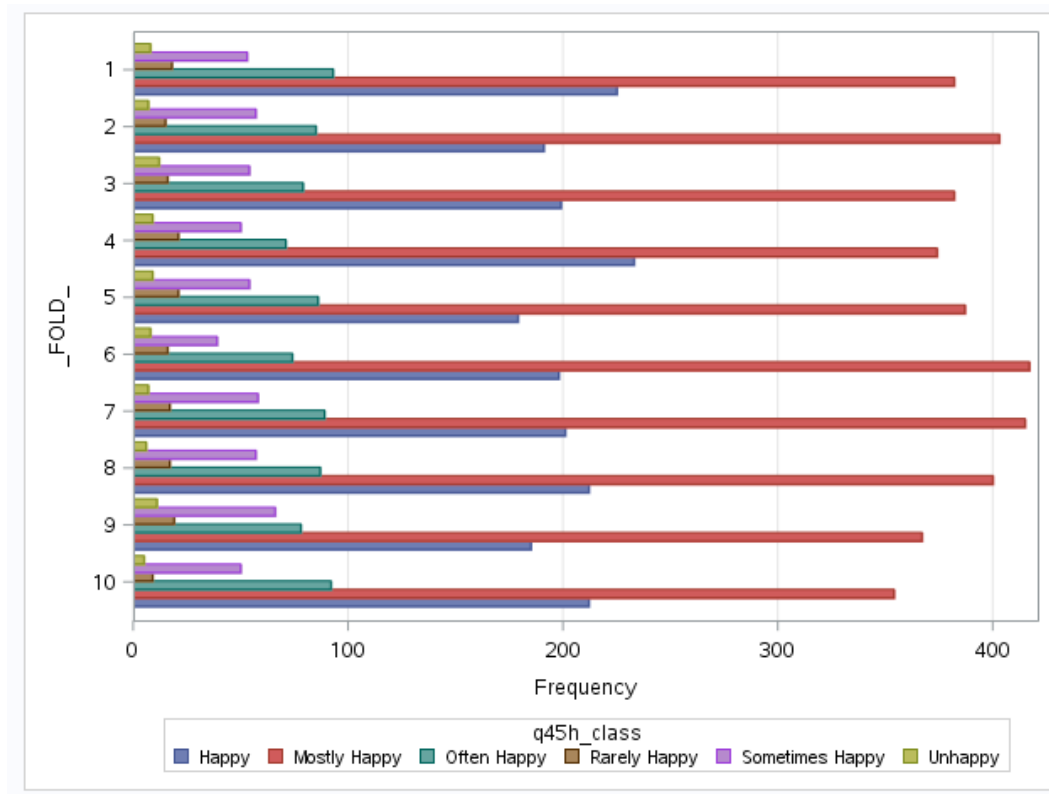


In case of both experiments, the SAS Enterprise Miner workflows were exactly the same and contained 11 nodes with first node being an Import of the source file (see Figure 4.4.)



**Figure 4.4 Experiment workflows**

The diagram in Figure 4.4 presents the implementation of stratified 10-fold validation resulting into 10 Misclassification Rate values being produced per model, which were then used to perform the testing for a statistical significance. In order to achieve this Transform Variables node was used to create a 10-fold cross validation indicator, which randomly divided dataset into 10-folds. This new variable (named ‘\_fold\_’) was setup as a segment variable, which is a requirement for cross-validation setup in the tool. Figure 4.5 presents the distribution of individual classes in each of the folds created.



**Figure 4.5 Distribution of target class in cross validation folds**

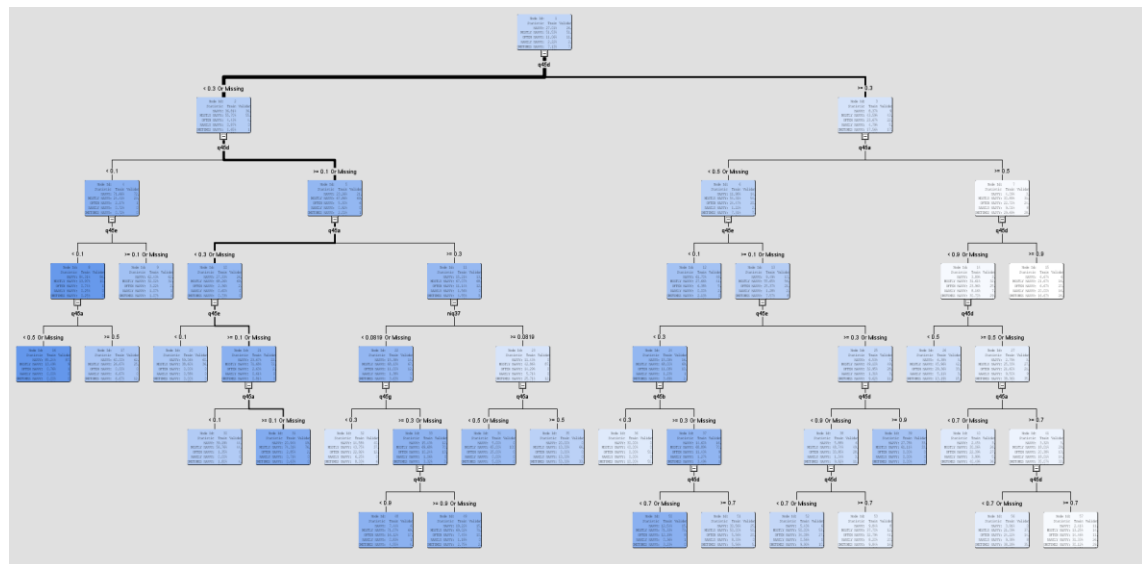
Afterwards, Start/End Groups nodes were implemented and their “Mode” was specified as “Cross-validation”. Start/End Group nodes are able to create 10 versions of training data and then calculate fit statistics of it, however they do not calculate over cross validation statistics. Thus, another node had to be used in order to obtain it: Model Import node. It is important to note, that due to the fact the Start and End Group nodes are used prior to Model Import Node and that the mode used is set to cross validation, all fit statistics produces are always listed as ‘Train:’. However, the ‘Train:’ part is actually the cross validation metric produced.

The last node used, the Model Comparison node, compares the fit statistics of a Decision Tree and Neural Networks models based on the 10-fold cross validation data. It provides the output table in which the training metrics are actually the 10-fold cross validation training and testing metrics, including the averages of Misclassification Rate, Average Squared Error and ROC index value.

#### 4.3.1. Decision Tree modelling

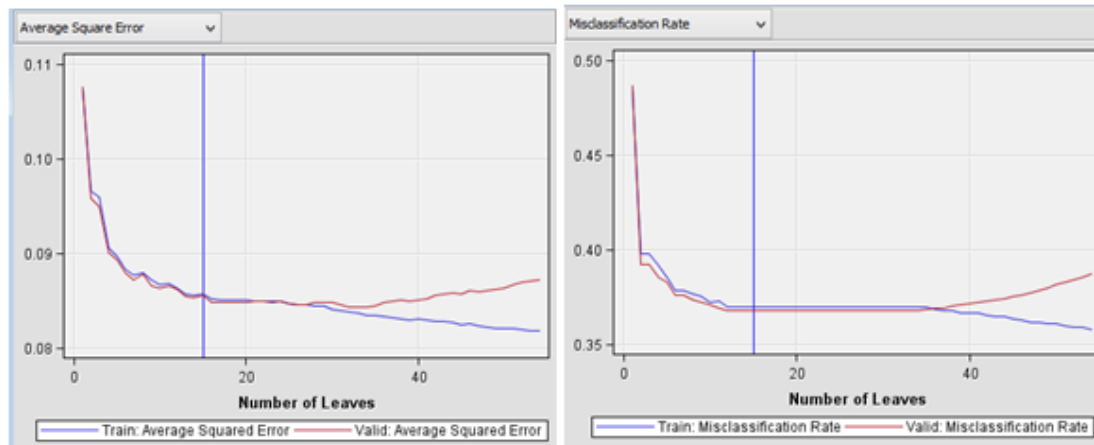
Before moving to the cross-validation and model comparison stage ten Decision Tree models were created and tested in order to determine the best settings for the Decision Tree. This step was performed using the full imbalanced dataset, which was split into 70% for training and 30% for validation parts.

First model created, shown in Figure 4.6, was setup using Average Square Error selection made on sub-tree feature of Assessment Measure. For this tree the analysis of Sub-tree Assessment Plots has shown that the majority of fit improvement is present in the first 5 splits, with the best validation performance for Misclassification Rate and Average Square Error metrics present from 12 to up to 23 leaves, which then slightly reduces.



**Figure 4.6 Initial Decision Tree created**

The above model creation was followed by additional parameters modification of tree setting including splitting rules criterion and node options manipulation. Most of the models created presented similar Average Square Error and Misclassification Rate metrics with the best performance on the validation starting at 12 leaves, after which it decreased, as the models became more complex. At the same time training set presented constant improvement (see figure 4.7).

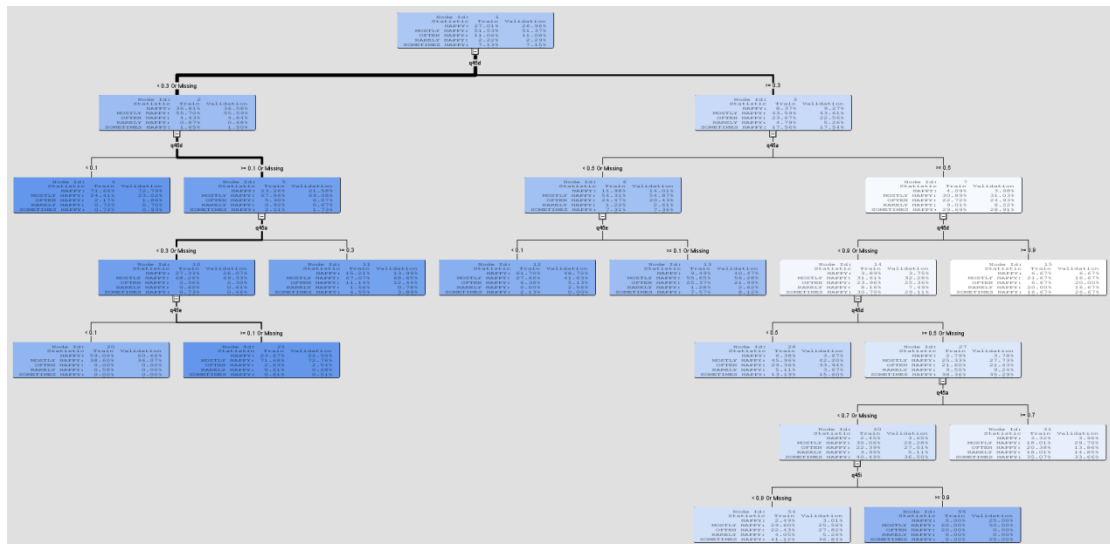


**Figure 4.7 Sub-tree Assessment Plots - over-fitting model**

The plots in Figure 4.7 show the Misclassification Rate and Average Squared Error corresponding to tree presented in Figure 4.6. For both of them the model performance on the training data becomes better as the tree becomes more complex. However, the performance on the validation only improves up to 18 leaves, and then decreases with model complexity. This type of performance difference between test and validation presents evidence of the model over-fitting, thus they were abandoned.

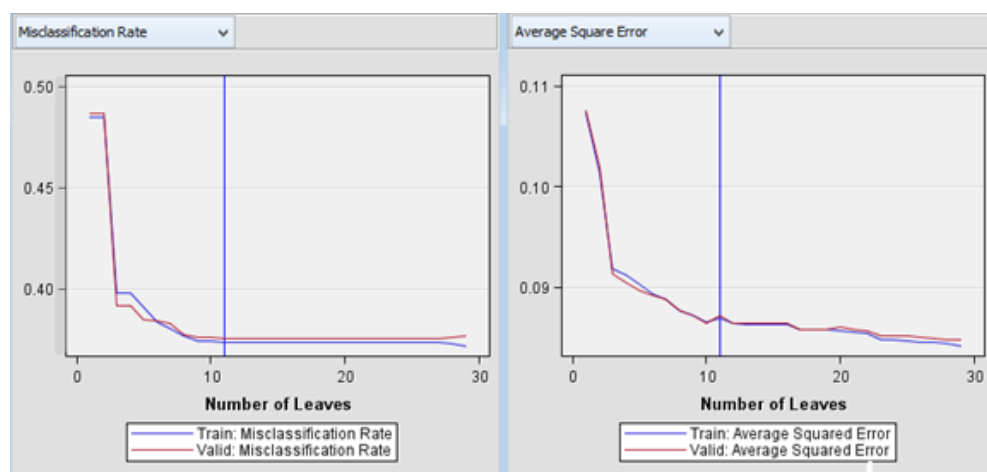
Finally, a Decision Tree selected for model comparison with Neural Networks was setup using Misclassification Rate sub-tree feature of Assessment Measure and was less deep than any other tree created (see Figure 4.8 below).

This resulted in the best achieved performance of both the Average Square Error and the Misclassification Rate fits. Similarly to the other models, both Assessment Metrics presented similar patterns of negative correlation being present between the values and the tree complexity for both: training and validation. However for this model the validation fit not only stayed optimal up to 27 splits, but also the grade of discrepancy present was significantly reduced. Table 4.7 present summary Fit Statistics for a Decision Tree model selected.



**Figure 4.8 Decision Tree – selected model settings**

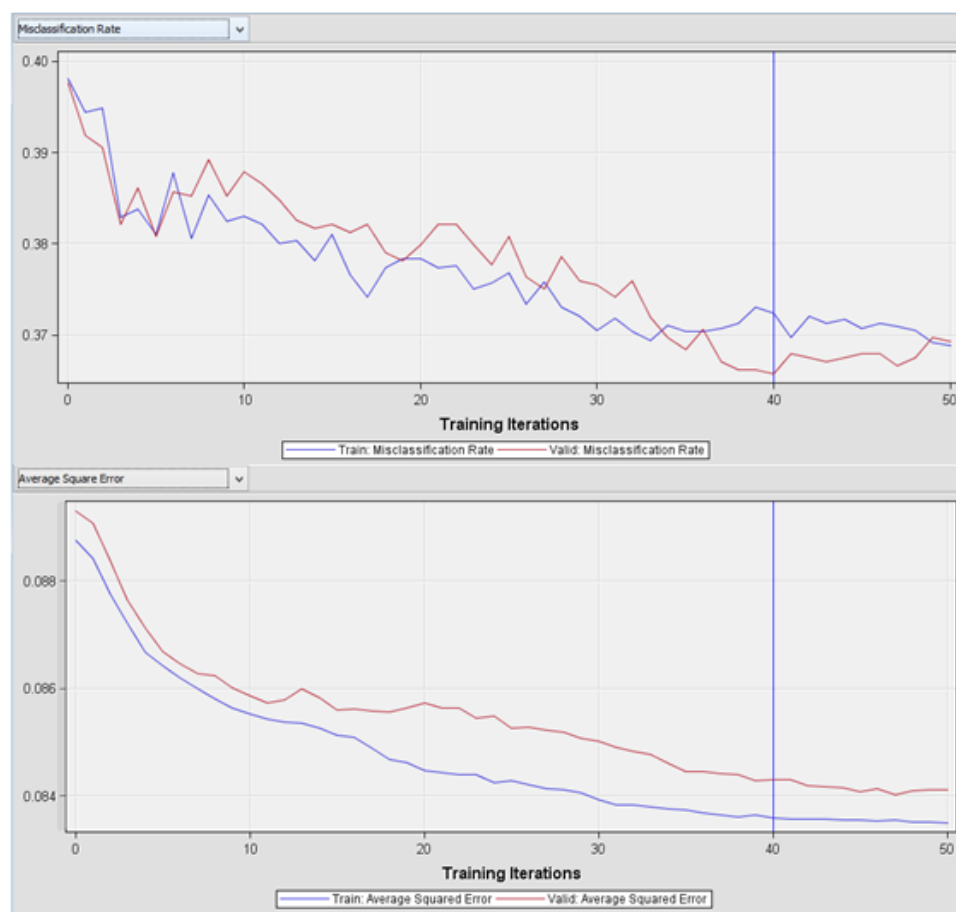
The plots in Figure 4.9 show the Misclassification Rate and Average Squared Error corresponding to final tree selected for model comparison. For both significant performance improvements on validation data (in comparison to the starting model) can be noticed. The split between the test and validation trend lines is reduced. This type of performance presents evidence of reduction in model over-fitting, thus the selection of model for final experiment step.



**Figure 4.9 Sub-tree Assessment Plots - model selected**

#### 4.3.2. Neural Networks modelling

Neural networks modelling step, just like it was for a Decision Tree model, involved creation of multiple models in order to find the one with the best performance. As previously discussed, Neural Networks model accuracy is strongly affected by the quantity of inputs. Thus, one of the models created and tested was setup to use Regression Model as a source of input variables. In this setup only the variables, which were selected by the regression (stepwise) weren't rejected. In addition, an AutoNeural model was created for performance comparison purposes. However, both of those models were excluded from the final comparison. The architecture of a Neural Networks model selected was modified and the number of hidden units was increased from 3 (default value) to 5, as any increase greater than 5 didn't have any significant effect on the value of the Average Squared Error or the Misclassification Rate.



**Figure 4.10 Iteration Plots - selected NN model**

As in case of Sub-tree Assessment plots analysed before, the plots in Figure 4.10 present the Misclassification Rate and Average Squared Error corresponding to final Neural Network model selected for performance comparison, however in case of this algorithm the plots analyse the fit statistics over the count of iterations. For both plots the test and validation performance improves as the amount of iterations increases, while the optimal point is marked at 40. The split between the test and validation is quite stable. There is no evidence of model over-fitting.

The summary of the Fit Statistics resulting from modelling step presented in the Table 4.7 shows that the performance of a Neural Network was slightly better than the one of a Decision Tree model. However, these figures required further evaluation and analysis using cross-validation techniques.

Model	Validation: Misclassification rate	Validation: Average Squared Error	Validation: ROC Index
Decision Tree	0.376	0.087	0.662
Neural Network	0.366	0.084	0.701

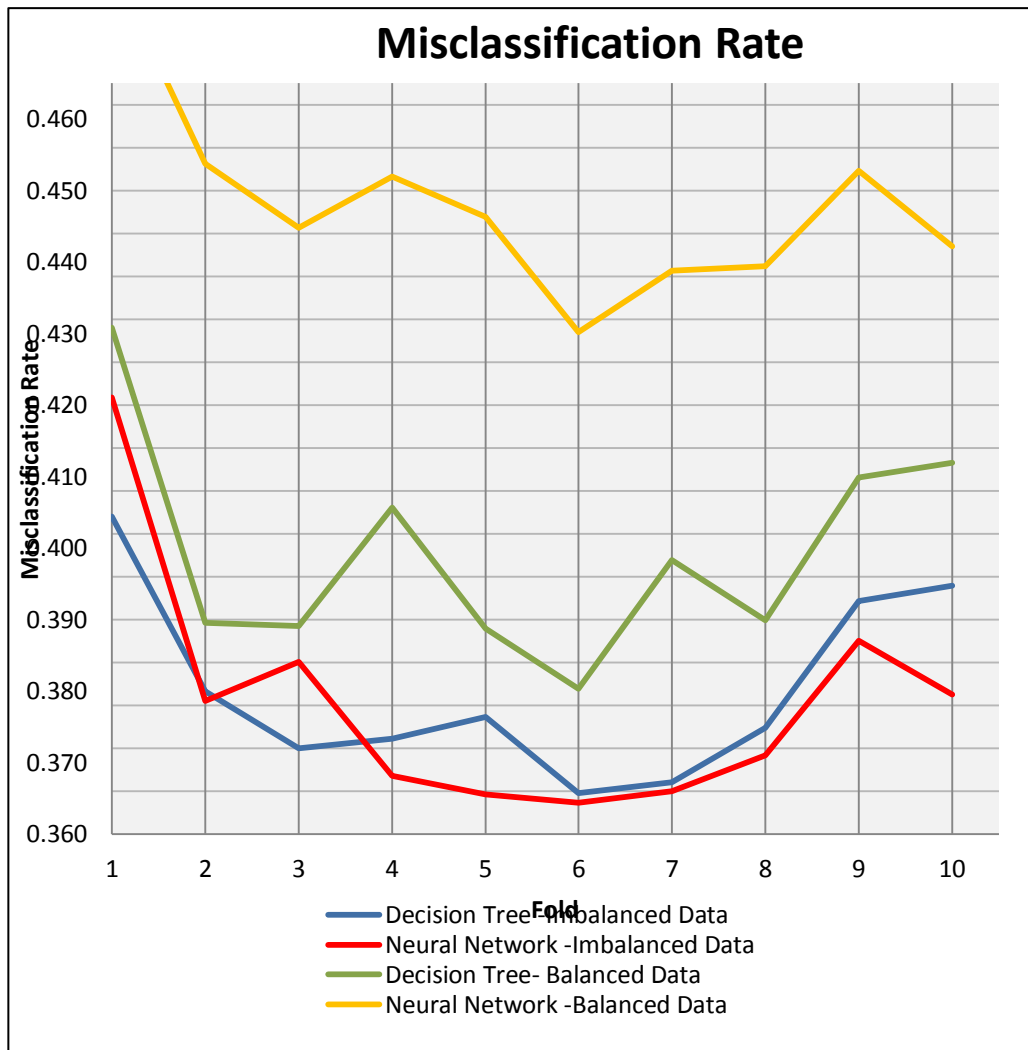
**Table 4.7 Fit Statistics - models selected**

#### **4.4. Evaluation**

This section of the reports presents the evaluation phase of the experiment, which includes both: the comparison of the model performance, as well as the testing for statistical significance.

##### **4.4.1. Model performance comparison**

Both experiments carried out used 2 classification models described in the modelling section of this chapter and 2 different data sources – the imbalanced and the balanced one. Since stratified 10-fold validation technique was used, not only an average of Misclassification Rate values were obtained from all, but also the values produced for each fold. The line chart below (Figure 4.11) presents the variation of Misclassification Rate for all models between different folds.



**Figure 4.11 Misclassification Rates of Model**

In the initial model comparison performed, using 70/30 training/validation split in modelling phase, Neural Networks model was presented as the better model - based on the Validation data: Misclassification rate and Average Squared Error comparisons.

Similarly, the results obtained in the first experiment, when models were run over the full dataset using 10-fold cross-validation using the original imbalanced data, Neural Networks fit statistics were also better than the ones of a Decision Tree model. However, it is important to note that the difference between the values decreased. The analysis of the ROC index values presented similar findings - showing slightly better performance of a Neural Networks model and an overall decrease in difference between the models.



However, the analysis of the experiment results (Misclassification Rate from 10-fold cross validation) for the balanced dataset had shown complete shift in the values, with the strong drop in performance of a Neural Network model resulting in a Decision Tree model being selected as the best fit. While a Decision Tree model presented the drop in performance of 0.019, a Neural Network model's Misclassification Rate increased by 0.069. Table 4.8 below presents the exact figures achieved by all model

Fold	Model			
	Decision Tree - Imbalanced Data	Neural Network - Imbalanced Data	Decision Tree- Balanced Data	Neural Network -Balanced Data
1	0.404	0.421	0.431	0.482
2	0.380	0.379	0.390	0.454
3	0.372	0.384	0.389	0.445
4	0.373	0.368	0.406	0.452
5	0.376	0.365	0.389	0.446
6	0.366	0.364	0.380	0.430
7	0.367	0.366	0.398	0.439
8	0.375	0.371	0.390	0.439
9	0.393	0.387	0.410	0.453
10	0.395	0.380	0.412	0.442
<b>Average</b>	<b>0.380</b>	<b>0.379</b>	<b>0.399</b>	<b>0.448</b>

**Table 4.8 Model Performance - Misclassification Rates**

At this stage it can't be said which one of the models should be considered as the more accurate on in terms of Misclassification rate comparison, as each of the experiments provided completely different results. Further analysis of the test results, including the testing for statistical significance in the performance difference is therefore required.

The confusion matrices presented in Figures 4.12 and 4.13 present all predictions made for the test by the models in both experiments. The results reported show that greatest loss in accuracy is present for the target class "3- A good bit of the time", as the models using the balanced data make almost no predictions belonging to this class. It also present the significant increase of overall predictions made by Neural Networks

model in classes:” 5 - A little of the time” (increase from 5 to 505) and “6 - None of the time” (increase from 9 to 523).

			Prediction								
			1 - All of the time	2 - Most of the time	3 - A good bit of the time	4 - Some of the time	5 - A little of the time	6 - None of the time			Total
Decision Tree	Target	1 - All of the time	1176	820	9	22	5	3	2035		TP
		2 - Most of the time	449	3187	37	192	9	7	3881		4674
		3 - A good bit of the time	37	605	25	148	12	7	834		Acc.
		4 - Some of the time	14	222	13	263	22	4	538		0.620
		5 - A little of the time	12	61	7	66	18	5	169		Misc
		6 - None of the time	13	31	1	19	13	5	82		Rate
		Total	1701	4926	92	710	79	31	7539		0.380
			Prediction								
			1 - All of the time	2 - Most of the time	3 - A good bit of the time	4 - Some of the time	5 - A little of the time	6 - None of the time			Total
Neural Network	Target	1 - All of the time	1149	855	13	17	1	0	2035		TP
		2 - Most of the time	434	3210	105	128	2	2	3881		4685
		3 - A good bit of the time	28	581	100	123	1	1	834		Acc.
		4 - Some of the time	13	240	62	223	0	0	538		0.621
		5 - A little of the time	13	70	12	71	0	3	169		Misc
		6 - None of the time	13	28	1	36	1	3	82		Rate
		Total	1650	4984	293	598	5	9	7539		0.379

Figure 4.12 Confusion Matrices for Imbalanced Data Experiment

			Prediction								
			1 - All of the time	2 - Most of the time	3 - A good bit of the time	4 - Some of the time	5 - A little of the time	6 - None of the time	Total		
Decision Tree	Target	1 - All of the time	1118	857	0	21	11	28	2035		TP
		2 - Most of the time	473	3134	1	207	28	39	3882		5552
		3 - A good bit of the time	36	597	0	163	22	16	834		Acc.
		4 - Some of the time	14	291	0	554	190	26	1075		0.601
		5 - A little of the time	12	180	0	174	448	31	845		Misc
		6 - None of the time	30	56	1	57	132	298	574		Rate
		Total	1683	5115	2	1176	831	438	9245		0.399
			Prediction								
			1 - All of the time	2 - Most of the time	3 - A good bit of the time	4 - Some of the time	5 - A little of the time	6 - None of the time	Total		
Neural Network	Target	1 - All of the time	1019	954	0	26	9	27	2035		TP
		2 - Most of the time	373	3157	0	223	65	63	3881		5101
		3 - A good bit of the time	24	543	0	192	54	21	834		Acc.
		4 - Some of the time	7	369	0	496	125	90	1087		0.552
		5 - A little of the time	21	264	0	284	186	79	834		Misc
		6 - None of the time	63	161	0	41	66	243	574		Rate
		Total	1507	5448	0	1262	505	523	9245		0.448

Figure 4.13 Confusion Matrices for Balanced Data Experiment

#### 4.4.2. Statistical significance and hypothesis evaluation

Final step of the research was to test a statistical significance of the results of the experiments performed. Thus, following hypotheses were tested in order to fully address the research question:

*H0: There is a statistically significant difference in the value of prediction accuracy of the subjective well –being between Neural Networks and Decision Trees with p-value <0.01*

*H1: There is no statistically significant difference in the value of prediction accuracy of the subjective well –being between Neural Networks and Decision Trees with p-value <0.01*

Paired t-test (Wilcoxon Signed-Rank) (Gibbons and Chakraborti, 2011) was performed on the Misclassification Rate values obtained from the 10-fold cross-validation of each model. The cut-off value determining the statistical significance chosen was 0.01.

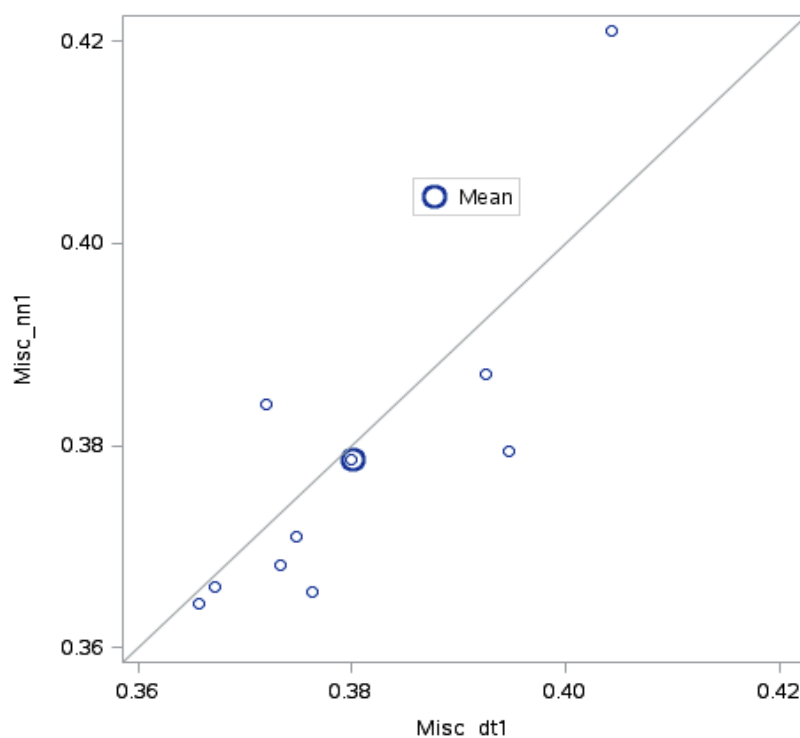
The test results achieved were different between the two experiments. For the first experiment the test has shown that the difference in the results is not statistically significant, with the p-value > 0.01. Thus, the result of the first experiment provided an evidence to reject the hypothesis H0 and accept H1

Experiment 1- Imbalanced Data				
Test	Statistic		p Value	
Student's t	t	0.522103	Pr >  t	0.6142
Sign	M	3	Pr >=  M	0.1094
Signed Rank	S	9.5	Pr >=  S	0.3750

**Table 4.9 Wilcoxon Signed-Rank Test Results - Experiment 1**

The test results presented in Table 4.9 can be additionally supported by the Agreement Plot (Figure 4.14) in which a Decision Tree Misclassification Rate results (Misc\_dt1) are plotted against a Neural Networks Misclassification Rates (Misc\_nn1). The Regression line has a slope of 1, and identifies the points where the difference is equal

to 0. In case of the first experiment's Agreement Plot, it is clearly visible that the majority of points, including the mean, are situated relatively closely to the regression line, which corresponds to the results of t-test.



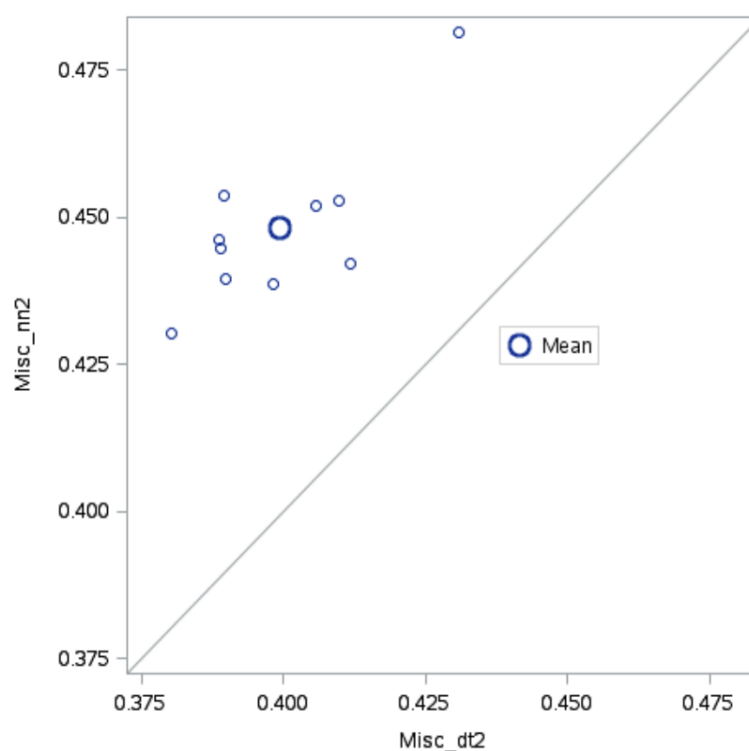
**Figure 4.14 Agreement Plot - Experiment 1**

Wilcoxon Signed-Rank t-Test results for the second experiment Misclassification Rates were different than the ones for the first experiment. In this case, the t-test has shown that the difference in the results of the models is statistically significant, with the p-value  $< 0.01$ . Therefore, the result of the second experiment provided an evidence to accept the hypothesis  $H_0$  and reject the  $H_1$ .

Experiment 2 - Balanced Data				
Test	Statistic		p Value	
Student's t	t	-16.1489	Pr >  t	<.0001
Sign	M	-5	Pr >=  M	0.0020
Signed Rank	S	-27.5	Pr >=  S	0.0020

**Table 4.10 Wilcoxon Signed-Rank Test Results - Experiment 2**

As it was for the first experiment, the test results presented in Table 4.10 can be also visualised using the Agreement Plot (Figure 4.15). In this plot a Decision Tree's Misclassification Rate is represented by variable 'Misc\_dt2' are plotted against a Neural Networks' Misclassification Rates (Misc\_nn2). As in previous graph, the regression line identifies the points where the difference is equal to 0. However, in case of Experiment 2 Agreement Plot the majority of points, including the mean, are situated relatively far from the regression line, which again corresponds to the results of t-test.



**Figure 4.15 Agreement Plot - Experiment 2**

All of the above does not provide definite answer to the research question. Analysis of all the experimental results, as well as Wilcoxon Signed-Rank t-Test results, allowed to do both: accept the hypothesis  $H_0$ , in case of Experiment 2, as well as reject  $H_0$ , in case of Experiment 1. As the results obtained are contradicting each other, it can't be concluded that the either of the models has a better performance as compared to the other one.

#### **4.5. *Experiment summary***

The main goal of the experiments performed was to collect the data regarding the performance of two techniques: Decision Tree and Neural Network classifier, and compare it, which will allow for identification of the outperforming model.

Prior experiment execution multiple strategies were developed for building data understanding and performing data pre-processing, all of which were studied in detail during the research and applied on the data when the evidence was found that they will improve the results.

The first experiment involved Decision Tree and Neural Network performance comparison using the dataset, which have undergone all the pre-processing changes except the imbalance removal using SMOTE. The second experiment also involved a Decision Tree and Neural Network performance comparison; however balanced dataset was used here instead.

The analysis of the result from the experiment one, where the imbalanced data was used, has shown slightly better performance of a Neural Networks model over the Decision Tree, however the 'Wilcoxon Signed-Rank Test' for statistical significance has shown that the difference in the results is not statistically significant, with the p-value  $> 0.01$ . Thus, first experiment provided an evidence to reject the  $H_0$  and accept  $H_1$ .

In case of the second experiment, where the balanced dataset was used, the analysis of the experiment results had shown strong drop in performance of a Neural Network model resulting in Decision Tree model being selected as the best fit. While a Decision Tree presented the increase of Misclassification Rate by 0.019, a Neural Network increased by 0.069. Additionally, the t-test has proved that the difference in the results is statistically significant, with the p-value  $< 0.01$ . Therefore, the result of the second experiment provided an evidence to accept the  $H_0$  and reject the  $H_1$ .

## 5. ANALYSIS AND DISCUSSION

This chapter provides a critical evaluation of strengths and limitations of the experiments implemented and described in the previous chapter. Discussion is made on different results obtained for two types of data used and their statistical significance. Further, the results are analysed in the context of past research, which was previously discussed in the literature review.

### 5.1. *Strength and limitations of results*

The design and methodology used for the implementation of the experiment took into consideration multiple issues, which could affect the results obtained in the final step.

The strategies developed for building data understanding and performing data pre-processing were well planned, and accounted for all the necessary activities. All the activities related to gaining data understanding, and performing data pre-processing were studied in detail during the research, and were applied on the data when the evidence was found that they can improve the results.

The goal of the research study was to investigate the performance of two Supervised Machine Learning models: Decision Tree and Neural network for the prediction of multiclass target variable. However, only during the feature selection step the awareness of very low correlation being present between all the independent variables and the target was gained. As correlation value informs how much information can be obtained from one variable regarding the other variable, the stronger the correlation, the easier it is to make predictions about one variable based upon another. Correlation values between the target and independent variables were not known at the time of hypotheses design, and in effect it allowed for achieving only 62.1% accuracy on all models tested. As overall model performance is affected by the correlation values present, it is suggested to verify those values at the beginning of any future work.

The modelling phase allowed determining the model parameters, which had significant impact on overall algorithms' accuracy, which is definitely the strength of this research. Moreover, comparison of models using stratified 10-fold validation allowed for obtaining not only average accuracy, but also the results allowing for testing for



statistical significance of the variance. However, performing the multiclass target classification using the “extension from binary” approach taken, where predictions for the multiclass classification problem are being made by extending some algorithms from the binary classification to multiclass, only allowed for achieving, previously mentioned accuracy of 62.1% (the best average accuracy achieved; Neural Networks Misclassification Rate = 0.379). Therefore, it would be suggested to perform other comparison in the future work, where the “reduction to binary” would be performed instead. Either of the approaches discussed in the literature review: one-versus-one (where each binary classifier is build using a pair of classes from the original data, and then final prediction is made using combined output from multiple binary classifiers) or one-versus-all (which involves creation and training of models per binary class, where one original class is reduced as positive, while all the other as negatives), could be selected for the purposes of new research.

Additionally, it was investigated if the imbalance present in the target variable affects the performance of the models, thus two experiments including two models of interest were conducted one using the imbalanced data and one using the balanced data (where SMOTE over-sampling technique was implemented). The results obtained here, showed that while the Decisions Tree accuracy remained on the approximately the same level (increase of Misclassification Rate from 0.38 to 0.399), the Neural Networks model was much more negatively affected with the Misclassification Rate value increasing from 0.379 to 0.448. This may indicate that using SMOTE on survey dataset might not always lead to overall improvement in classification, and that each case should be always examined by using both: the original source and the modified version of it.

The results of ‘Wilcoxon Signed-Rank Test’ for statistical significance were different between the two experiments. For the first experiment, where the original imbalanced data was used, the test has shown that the difference in the results is not statistically significant, with the p-value  $> 0.01$ . Thus, the result of the first experiment provided an evidence to reject the hypothesis  $H_0$  and accept  $H_1$ . In case of the second experiment, where the balanced dataset was used, the t-test has shown that the difference in the results of the models is statistically significant, with the p-value  $< 0.01$ . Therefore, the result of the second experiment provided an evidence to accept the hypothesis  $H_0$  and reject the  $H_1$ . However, this experiment also presented negative effect on the model

performance, which may suggest that imbalance removal using SMOTE, where new minority class instances are created by interpolating between several examples from minority class, is not suitable when working with survey data, where numeric values are only discrete.

To conclude, both the strength and the limitations of the results focus on the data pre-processing techniques selected in order to improve the performance of the models. It is possible that different results would be obtained if different approach was selected for multiclass target and/or different imbalance removal technique was used.

## **5.2. *Considerations in regards to previous research***

Literature review performed in regards to the research on the subjective well-being provided the evidence that the concept is affected by a number of separable, although related, factors. Modern research, including studies conducted by psychologists, sociologists and economists, increased the understanding of how the individual components (or factors) affect the subjective well-being. The main groups of factors include economic circumstances, social relationships, as well as health and health related factors. However, as the purpose of this research was to verify the use of Machine Learning algorithms in the prediction of SWB using survey data, which includes questions related to the different groups of factors, feature selection (using Pearson's correlation coefficient) had to be performed prior to modelling.

Analysis of the results produced by the algorithm has shown, that although all of the features selected can be grouped by using groups indicated in past research (i.e. General Health, Mental Health, Diet and Nutrition, Physical Activity and Social Connectedness), all of them present really low level of correlation value with the target variable with the range from 0.0296 to 0.1822. This stands in opposition to the some of the previous research where strong correlation was identified, e.g. Fernández-Ballesteros, *et.al.* (2001), Gerlach and Stephan (1996), Dolan, Peasgood, and White (2008). However, it is possible that this difference may be a result of the survey structure and further investigation could be conducted, which would compare the format of questions and structure of data between different experiments. If any significant differences were identified they could be used to modify the current Healthy Ireland Survey format in the future, and in effect improve the data collection.

This could lead to the overall improvement in addressing common issues, and in effect the increase of Irish population well-being.

In regards to Machine Learning aspect of this research, two main points of interest can be highlighted: first is the performance of Supervised Machine Learning classification algorithms (Decision Tree and Neural Network) on the multiclass target using extension from binary approach, and the second is the impact of the imbalance presence on the performance of those algorithms.

As previously discussed, performing the classification using the extension from binary approach taken allowed for achieving only 62.1% accuracy (the best average accuracy achieved; Neural Networks Misclassification Rate = 0.379). Literature review performed, in particular research by Aly (2005), provided evidence that both Decision Trees and Neural Networks can solve the multiclass classification problem by extending the binary classification technique. However, achieved level of accuracy indicates that the other approaches, where multiclass classification problem is converted into a set of binary problems, should be investigated as well, as they may improve the overall accuracy of the models. In effect it could be proven, that classification algorithms present better results for making prediction based on survey data, when using different approach to multiclass classification problem.

Finally, while the review of research by Weiss and Provost (2001), and Chawla, Japkowicz, and Kotcz (2004), He and Garcia (2009) Batista, *et.al.* (2005) led to selection of SMOTE for the imbalance reduction, as the authors documented that for multiple base classifiers, imbalance removal led to overall improvement in classification performance and claimed that SMOTE it the best method which doesn't risk information loss, or over-fitting of models created, the results achieved from the experiment are contradictory. The Misclassification Rate results produced by models using balanced data were higher than the ones where the original, imbalanced data was used. This may suggest that using SMOTE on survey dataset might not always lead to the overall improvement in classification, and that each case should be always examined by using both: the original source and the modified version of it.

## **6. CONCLUSION**

### **6.1. *Research Overview***

The purpose of this research was to investigate and compare the performance of two supervised machine learning techniques for the prediction of a multiclass target, where imbalance is present. It would extend existing research on application of Machine Learning in the area of SWB. The ability to make predictions regarding one's SWB could be valuable, for example, in relation to identification of other possible negative outcomes resulting from low subjective well-being e.g. suicide, depression, etc.

The main goal of the study was to collect the data regarding the performance of two techniques, Decision Tree and Neural Network classifier, and compare it, which will allow for identification of the outperforming model.

Two experiments using above supervised classification techniques were conducted. The first experiment involved Decision Tree and Neural Network performance comparison using a dataset which had undergone all the pre-processing changes except imbalance removal using SMOTE. The second experiment also involved Decision Tree and Neural Network performance comparison; however a balanced dataset was used here instead.

Decision Trees were selected as they are the most fundamental machine learning models, which are able to provide interpretability and information about the importance of individual features. Additionally, they were identified in the Literature Review as the benchmark algorithm to which other supervised learning algorithms should be compared. At the same time, investigation of Neural Networks has proven their suitability and previous high performance on multiclass classification problems.

### **6.2. *Problem Definition***

The literature review conducted provided an overview of most important and state-of-art research related to both subjective well-being and Machine Learning, and the gaps

and limitations identified through it provided motivation for the following research question definition:

*Which of the classifiers, Decision Trees or Neural Networks, is more accurate in predicting subjective 'well-being' with the use of specified economic, social and health related factors?*

Therefore, the following hypotheses were considered to allow for addressing above research question:

*H0: There is a statistically significant difference in the value of prediction accuracy of the subjective well-being between Neural Networks and Decision Trees with  $p\text{-value} < 0.01$*

*H1: There is no statistically significant difference in the value of prediction accuracy of the subjective well-being between Neural Networks and Decision Trees with  $p\text{-value} < 0.01$*

In order to achieve answer to the above question an experiment was conducted. Selection of accurate methodology was crucial to the process, as it allowed for addressing any data issues identified, implementation of any required data pre-processing solutions, and the achievement of the best performance.

### **6.3. Design/Experimentation, Evaluation & Results**

The analysis of the result from the first experiment, where imbalanced data was used, showed a slightly better performance of a Neural Networks model over the Decision Tree, however the 'Wilcoxon Signed-Rank Test' for statistical significance has shown that the difference in the results is not statistically significant, with the  $p\text{-value} > 0.01$ . Thus, the first experiment provided an evidence to reject the H0 and accept H1.

In case of the second experiment, where the balanced dataset was used, the analysis of the experiment results had shown a strong drop in performance of a Neural Network model resulting in Decision Tree model being selected as the best fit. While a Decision Tree presented the increase of Misclassification Rate by 0.019, a Neural Network increased by 0.069.

Additionally, the t-test has proved that the difference in the results is statistically significant, with the  $p\text{-value} < 0.01$ . Therefore, the result of the second experiment

provided evidence to accept the H0 and reject the H1. As the results obtained in both experiments are contradictory, it would be suggested to test the performance of other supervised machine learning models in order to the data usability for the prediction making.

#### **6.4. *Contributions and impact***

This research explored the application of classification algorithms for the prediction of the self-reported value of subjective well-being. The experiment conducted resulted in identification of multiple findings, not only related to the performance of models itself, but also to the impact of the strategies and approaches taken on the value of model performance, those include mainly:

- Although all of the features selected for model building can be grouped by using factor groups indicated in research (i.e. General and Mental Health, Diet and Nutrition, Physical Activity and Social Connectedness), all of them present really low levels of positive correlation value with the target variable with the range from 0.0296 to 0.1822. This stands in opposition to the some of the previous research discussed in literature review, where strong correlation was identified.
- There is no statistically significant difference between the performance of Decision Tree and Neural Network, when performing the classification using the extension from binary approach.
- Imbalance removal using SMOTE had a negative effect on the model performance, which may suggest that this approach, which creates new minority class instances by interpolating between several examples from minority class, is not suitable when working with survey data, where numeric values are only discrete.

### **6.5. *Future Work & recommendations***

As this project only focused on two algorithms, Decision Tree and Neural Networks, further research in regards to performance comparison of such models as k-Nearest Neighbour, Naive Bayes, and Support Vector Machines is required

Moreover, future work could be done to verify the levels of the prediction accuracy for models used in this research, however different approaches could be selected for handling the multiclass target classification. As discussed in the literature review, the reduction to binary is another common approach used for handling to multiclass classification problem. Therefore, it would be of value to compare the results from this research to research using reduction to binary.

Finally, it would be suggested to attempt to design the survey with the machine learning experiment in mind, where the question and the structure are more compatible with machine learning and predictive models creation. Different designs could be tested and compared in order to verify the most effective structure, which could then be used as a guideline and/or recommended template for any future nationwide Health Related (including subjective well-being) surveys conducted, not only in Ireland, but in other countries.

## BIBLIOGRAPHY

Adler, M. D., Dolan, P., & Kavetsos, G. (2017). Would you choose to be happy? Trade-offs between happiness and the other dimensions of life in a large population survey. *Journal of Economic Behaviour & Organization*, 139, 60-73. doi:10.1016/j.jebo.2017.05.006

Allwein, E. L., Schapire, R. & Singer, Y. (2000), Reducing multiclass to binary: A unifying approach for margin classifiers, *Journal of Machine Learning Research* 1, 113-141. Retrieved from: <http://www.jmlr.org/papers/volume1/allwein00a/allwein00a.pdf>

Aly, M. (2005). Survey on multiclass classification methods. *Neural Networks Technical Report, Caltech.*, 19, 1-9., doi:10.1.1.175.107

Baker, L. A., Cahalin, L. P., Gerst, K., & Burr, J. A. (2005). Productive activities and subjective well-being among older adults: The influence of number of activities and time commitment. *Social Indicators Research*, 73(3), 431-458. doi: 10.1007/s11205-005-0805-6

Batista, G. E., Prati, R. C., & Monard, M. C. (2005, September). Balancing strategies and class overlapping. In *International Symposium on Intelligent Data Analysis* (pp. 24-35). Springer, Berlin, Heidelberg. Retrieved from: <http://conteudo.icmc.usp.br/pessoas/gbatista/files/ida2005.pdf>

Becchetti, L., & Rossetti, F. (2009). When money does not buy happiness: The case of “frustrated achievers”. *The Journal of Socio-Economics*, 38(1), 159-167. doi:10.1016/j.socec.2008.08.009

Benjamin, D. J., Kimball, M. S., Heffetz, O., & Szembrot, N. (2014). Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference. *The American Economic Review*, 104(9), 2698–2735. <http://doi.org/10.1257/aer.104.9.2698>



- Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, 5(Sep), 1089-1105., Retrieved from <http://www.jmlr.org/papers/v5/grandvalet04a.html?92f58540>
- Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7), 1145-1159., doi: [https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2)
- Breiman, L. (2017). *Classification and regression trees*. Routledge. Pp.2016-2064
- Carbonell, J. G., Michalski, R. S., & Mitchell, T. M. (1983). *An Overview Of Machine Learning*. Machine Learning, Elsevier Inc., pp. 3-23. doi:10.1016/b978-0-08-051054-5.50005-4
- Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on Machine learning - ICML 06*. doi:10.1145/1143844.1143865
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide* (). The CRISP-DM consortium. Retrieved from: <http://www.crisp-dm.org/CRISPWP-0800.pdf>
- Chavan, G. S., Manjare, S., Hegde, P., & Sankhe, A. (2014). A Survey of Various Machine Learning Techniques for Text Classification. *International Journal of Engineering Trends and Technology*, 15(6), 288-292. doi: 10.14445/22315381/ijett-v15p25
- Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial. *ACM SIGKDD Explorations Newsletter*, 6(1), 1. doi:10.1145/1007730.1007733
- Conry, M. C., Morgan, K., Curry, P., McGee, H., Harrington, J., Ward, M., & Shelley, E. (2011). The clustering of health behaviours in Ireland and their relationship with mental health, self-rated health and quality of life. *BMC Public Health*, 11(1). doi:10.1186/1471-2458-11-692
- Costa, E., Lorena, A., Carvalho, A. C. P. L. F., & Freitas, A. (2007). A review of performance evaluation measures for hierarchical classifiers. In *Evaluation Methods*

*for machine Learning II: papers from the AAAI-2007 Workshop* (pp. 1-6). Retrieved from: <http://www.aaai.org/Papers/Workshops/2007/WS-07-05/WS07-05-001.pdf>

Coyle, L. D., & Vera, E. M. (2013). Uncontrollable stress, coping, and subjective well-being in urban adolescents. *Journal of Youth Studies*, 16(3), 391-403.doi: 10.1080/13676261.2012.756975

Department of Health. (2016), Health and Wellbeing Programme. Healthy Ireland Survey, 2015 [computer file]. Dublin: Irish Social Science Data Archive [distributor], March 2016

Daniely, A., Sabato, S., Ben-David, S., & Shalev-Shwartz, S. (2011). Multiclass learnability and the erm principle. In *Proceedings of the 24th Annual Conference on Learning Theory* (pp. 207-232). Retrieved from: <http://proceedings.mlr.press/v19/daniely11a.html>

Daniely, A., Sabato, S., & Shwartz, S. S. (2012). Multiclass learning approaches: A theoretical comparison with implications. In *Advances in Neural Information Processing Systems* (pp. 485-493). Retrieved from: <http://papers.nips.cc/paper/4678-multiclass-learning-approaches-a-theoretical-comparison-with-implications>

Dietterich, T. G. & Bakiri, G. (1995), 'Solving multiclass learning problems via error correcting output codes', *Journal of Artificial Intelligence Research* 2, 263-286. Retrieved from: <https://www.jair.org/media/105/live-105-1426-jair.pdf>

Diener, E., Lucas, R. E. & Oishi, S. (2002). Subjective well-being: The science of happiness and life satisfaction. In C.R. Snyder & S.J. Lopez (Eds.), *The handbook of positive psychology* (pp. 63-73). New York, NY:Oxford University Press.

Diener, E. (1984). Subjective well-being. *Psychological bulletin*, 95(3), 542. Retrieved from: [https://internal.psychology.illinois.edu/~ediener/Documents/Diener\\_1984.pdf](https://internal.psychology.illinois.edu/~ediener/Documents/Diener_1984.pdf)

Diener, E., & Suh, E. (1997). Measuring quality of life: Economic, social, and subjective indicators. *Social indicators research*, 40(1-2), 189-216.doi: 10.1023/A:1006859511756

Dolan, P., Peasgood, T., & White, M. (2008). Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. *Journal of Economic Psychology*, 29(1), 94-122. doi: 10.1016/j.joep.2007.09.001

Eurostat, (2015). *Quality of life Facts and views*. Luxembourg: Publications Office of the European Union, doi:10.2785/59737

Eurostat, (2017), *Final report of the expert group on quality of life indicators*. Luxembourg: Publications Office of the European Union, doi: 10.2785/021270

Fernández-Ballesteros, R., Zamarrón, M. D., & Ruíz, M. A. (2001). The contribution of socio-demographic and psychosocial factors to life satisfaction. *Ageing and Society*, 21(01), 25-43. doi:10.1017/s0144686x01008078

Ferrer-I-Carbonell, A. (2005). Income and well-being: an empirical analysis of the comparison income effect. *Journal of Public Economics*, 89(5-6), 997-1019. doi: 10.1016/j.jpubeco.2004.06.003

Finlay, S. (2014). *Predictive analytics, data mining and big data: myths, misconceptions and methods*. Basingstoke: Palgrave Macmillan., pp. 6 - 9

Fox, K. R. (1999). The influence of physical activity on mental well-being. *Public health nutrition*, 2(3a), 411-418. <https://doi.org/10.1017/S1368980099000567>

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L. (2009). Weka-a machine learning workbench for data mining. In *Data mining and knowledge discovery handbook* (pp. 1269-1277). Springer, Boston, MA. DOI: 10.1007/978-0-387-09823-4\_66

Frey, B. S., & Stutzer, A. (2005). Testing Theories of Happiness. *Economics and Happiness: Framing the Analysis*, 116-146. doi: 10.1093/0199286280.003.0005

Gerdtham, U., & Johannesson, M. (2001). The relationship between happiness, health, and socio-economic factors: results based on Swedish microdata. *The Journal of Socio-Economics*, 30(6), 553-557. doi:10.1016/s1053-5357(01)00118-4

- Gerlach, K., & Stephan, G. (1996). A paper on unhappiness and unemployment in Germany. *Economics Letters*, 52(3), 325-330. doi: 10.1016/S0165-1765(96)00858-0
- Gibbons, J. D., & Chakraborti, S. (2011). Nonparametric statistical inference. In *International encyclopedia of statistical science* (pp. 977-979). Springer Berlin Heidelberg. doi: [https://doi.org/10.1007/978-3-642-04898-2\\_420](https://doi.org/10.1007/978-3-642-04898-2_420)
- Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, 25(2), 141-151. doi:10.11613/bm.2015.015
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), 1157-1182. Retrieved from: [https://pdfs.semanticscholar.org/8300/82629772e85e8f1432fc12e54dfd9cfa4abd.pdf?\\_ga=2.109512965.120459059.1521480628-578470256.1521480628](https://pdfs.semanticscholar.org/8300/82629772e85e8f1432fc12e54dfd9cfa4abd.pdf?_ga=2.109512965.120459059.1521480628-578470256.1521480628)
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1), 10-18. 10.1145/1656274.1656278
- Hall, P., Dean, J., Kabul, I. K., & Silva, J. (2014). An overview of machine learning with SAS® enterprise miner™. White Paper SAS313-2014, *SAS Institute Inc.*, Retrieved from: <https://support.sas.com/resources/papers/proceedings14/SAS313-2014.pdf>
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier. ISBN 13: 978-1-55860-901-3, pp.291-309, 327-336
- He, H. and Garcia E. A. (2009), Learning from Imbalanced Data, in *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, Sept. 2009. doi: 10.1109/TKDE.2008.239
- Hunt, E. B., Marin, J., & Stone, P. J. (1966). Experiments in induction. Oxford, England: Academic Press.
- Iniesta, R., Stahl, D., & McGuffin, P. (2016) . Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, 46(12), 2455-2465. doi: 10.1017/S0033291716001367

Ipsos, M. R. B. I. (2016). *Healthy Ireland survey 2015: summary of findings*. Department of Health. Retrieved from: <http://www.healthyireland.ie/accessibility/healthy-ireland-survey/>

Jaques, N., Taylor, S., Nosakhare, E., Sano, A., Picard, R., (2016), Multi-task Learning for Predicting Health, Stress, and Happiness. In Proc. NIPS Workshop on Machine Learning in Health, Barcelona, Spain, December 2016. Retrieved from: <https://www.media.mit.edu/publications/multi-task-learning-for-predicting-health-stress-and-happiness/>

Kaplan, S. (2017). Deep Generative Models for Synthetic Retinal Image Generation. Retrieved from: [https://www.researchgate.net/publication/319093376\\_DEEP\\_GENERATIVE\\_MODELS\\_FOR\\_SYNTHETIC\\_RETINAL\\_IMAGE\\_GENERATION](https://www.researchgate.net/publication/319093376_DEEP_GENERATIVE_MODELS_FOR_SYNTHETIC_RETINAL_IMAGE_GENERATION)

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2), 271-277. Retrieved from: <https://pdfs.semanticscholar.org/3555/1bc9ec8b6ee3c97c524f9c9ceee798c2026e.pdf>

Mitchell, T. M. (2006). *The discipline of machine learning* (Vol. 9). Carnegie Mellon University, School of Computer Science, Machine Learning Department. Chicago, Retrieved from: <http://www-cgi.cs.cmu.edu/~tom/pubs/MachineLearningTR.pdf>

Mentzakis, E., & Moro, M. (2009). The poor, the rich and the happy: Exploring the link between income and subjective well-being. *The Journal of Socio-Economics*, 38(1), 147-158. doi:10.1016/j.socec.2008.07.010

Moran, S., He, Y., & Liu, K. (2009). Choosing the best Bayesian classifier: An empirical study. *IAENG International Journal of Computer Science*, 36(4), 322-331. Retrieved from: [https://www.researchgate.net/profile/Kecheng\\_Liu/publication/40422668\\_Choosing\\_the\\_Best\\_Bayesian\\_Classifier\\_An\\_Empirical\\_Study/links/56a269c108aef91c8c0eec16.pdf](https://www.researchgate.net/profile/Kecheng_Liu/publication/40422668_Choosing_the_Best_Bayesian_Classifier_An_Empirical_Study/links/56a269c108aef91c8c0eec16.pdf)

Moreno-Torres, J. G., Sáez, J. A., & Herrera, F. (2012). Study on the impact of partition-induced dataset shift on  $k$ -fold cross-validation. *IEEE Transactions on*

*Neural Networks and Learning Systems*, 23(8), 1304-1312.  
doi: 10.1109/TNNLS.2012.2199516

Nilsson, N. J., Sejnowski, T. J., & White, H. (1965). *Learning machines*. New York: McGraw-Hill Book Company. Retrieved from: <http://ia800806.us.archive.org/13/items/LearningMachines/Learning%20Machines.pdf>

North, R. J., Holahan, C. J., Moos, R. H., & Cronkite, R. C. (2008). Family support, family income, and happiness: A 10-year perspective. *Journal of Family Psychology*, 22(3), 475-483. DOI: 10.1037/0893-3200.22.3.475

Oswald, A J. and Powdthavee, N. (2008) *Does happiness adapt? : a longitudinal study of disability with implications for economists and judges*. *Journal of Public Economics*, Vol.92 (No.5/6). pp. 1061-1077. doi:10.1016/j.jpubeco.2008.01.002

Pal, S., & Mitra, S. (1992). Multilayer perceptron, fuzzy sets, and classification. *IEEE Transactions on Neural Networks*, 3(5), 683-697. doi:10.1109/72.159058

Pedersen, P. J., & Schmidt, T. D. (2011). Happiness in Europe. *The Journal of Socio-Economics*, 40(5), 480-489. doi:10.1016/j.socec. 2010.10.004

Pelckmans, K., Brabanter, J. D., Suykens, J., & Moor, B. D. (2005). Handling missing values in support vector machine classifiers. *Neural Networks*, 18(5-6), 684-692. doi:10.1016/j.neunet.2005.06.025

Pozzolo D. (2016) Racing for unbalanced method selection -Presentation of the unbalanced R package. Slideshare.net. (Feb 24, 2016). Retrieved from: <https://www.slideshare.net/dalpozz/presentation-of-the-unbalanced-r-package>

Quinlan, J. R. (1986). Induction of Decision Trees. *Mach. Learn.* 1, 1 (Mar. 1986), 81–106, DOI: 10.1007/BF00116251

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. pp. 27-32

Refaeilzadeh P., Tang L., Liu H. (2009) Cross-Validation. In: LIU L., ÖZSU M.T. (eds) *Encyclopedia of Database Systems*. Springer, Boston, MA. (pp. 532-538). Springer US. DOI: 10.1007/978-0-387-39940-9\_565

Rey D., Neuhäuser M. (2011) Wilcoxon-Signed-Rank Test. In: Lovric M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg

Rocha, A., & Goldenstein, S. K. (2014). Multiclass from binary: Expanding one-versus-all, one-versus-one and ecoc-based approaches. *IEEE Transactions on Neural Networks and Learning Systems*, 25(2), 289-302., doi: 10.1109/TNNLS.2013.2274735

Rosenblatt, F. (1961). *Principles of neurodynamics. perceptrons and the theory of brain mechanisms* (No. VG-1196-G-8). CORNELL AERONAUTICAL LAB INC BUFFALO NY. Retrieved from: <http://www.dtic.mil/dtic/tr/fulltext/u2/256582.pdf>

Schoon, I., Hansson, L., & Salmela-Aro, K. (2005). Combining Work and Family Life. *European Psychologist*, 10(4), 309-319. doi:10.1027/ 1016-9040.10.4.309

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press. Retrieved from: <http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning>

Shalabi L. A. and Shaaban Z. (2006), "Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix," *2006 International Conference on Dependability of Computer Systems*, Szklarska Poreba, 2006, pp. 207-214. doi: 10.1109/DEPCOS-RELCOMEX.2006.38

Shields, M., & Wheatley Price, S. (2005). Exploring the economic and social determinants of psychological wellbeing and perceived social support in England. *Journal Royal Statistical Society*(Part 3), 513–537. Doi: 10.1111/j.1467-985X.2005.00361.x

Siegel, E. (2016). *Predictive analytics: the power to predict who will click, buy, lie, or die*. Hoboken, NJ: Wiley., pp.1-16.

Silipo, R., Adae, I., & Hart, A. (2015). Seven techniques for data dimensionality reduction. Retrieved from: [https://mineracaodedados.files.wordpress.com/2015/06/knime\\_seventechniquesdatadimreduction.pdf](https://mineracaodedados.files.wordpress.com/2015/06/knime_seventechniquesdatadimreduction.pdf)

Stoica, C. (2015). Sleep, a Predictor of Subjective Well-being. *Procedia - Social and Behavioural Sciences*, 187, 443-447. doi:10.1016/j.sbspro. 2015.03.083

Swain, P. H., & Hauska, H. (1977). A Decision Tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3), 142-147. , doi: 10.1109/TGE.1977.6498972

Tatarkiewicz, W. (1976). Analysis of happiness. *Philosophy and Phenomenological Research* 38 (1):139-140, doi: 10.2307/2106529

Umberson, D., & Montez, J. K. (2010). Social Relationships and Health: A Flashpoint for Health Policy. *Journal of Health and Social Behaviour*, 51(1\_suppl): S54-S66. doi:10.1177/ 0022146510383501

Weiss, G. M., & Provost, F. (2001). The Effect of Class Distribution on Classifier Learning: An Empirical Study. *Technical Report MLTR-43*, Dept. of Computer Science, Rutgers University. Retrieved from: [https://www.researchgate.net/publication/2364670\\_The\\_Effect\\_of\\_Class\\_Distribution\\_on\\_Classifier\\_Learning\\_An\\_Empirical\\_Study](https://www.researchgate.net/publication/2364670_The_Effect_of_Class_Distribution_on_Classifier_Learning_An_Empirical_Study)

WHO, (2005), Promoting mental health: concepts, emerging evidence, practice. Geneva, World Health Organization. Retrieved from: [http://www.who.int/mental\\_health/publications/promoting\\_mh\\_2005/en/](http://www.who.int/mental_health/publications/promoting_mh_2005/en/)

Zhang, D., & Lee, W. S. (2003). Question classification using support vector machines. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR 03. doi:10.1145/860435.860443

Zoonen, W. V., & Toni, G. V. (2016). Social media research: The application of supervised machine learning in organizational communication research. *Computers in Human Behaviour*, 63, 132-141. doi:10.1016/j.chb.2016.05.028



## APPENDIX A: SAS CODE

```

libname source '/folders/myfolders';

/*creating nominal class using coded values*/
data hidata; set source.hidata ;
length q45h_class $30.;
if q45h = 1 then q45h_class = '1 - All of the time';
if q45h = 2 then q45h_class = '2 - Most of the time';
if q45h = 3 then q45h_class = '3 - A good bit of the time';
if q45h = 4 then q45h_class = '4 - Some of the time';
if q45h = 5 then q45h_class = '5 - A little of the time';
if q45h = 6 then q45h_class = '6 - None of the time';
label
q45h_class = 'How much time during the past 4 weeks..Have you been a happy person?';
run;

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=hidata;
    title height=14pt "Target Variable Distribution";
    vbar q45h_class / datalabel;
    yaxis grid;
run;

ods graphics / reset;
title;

/*summary of missing values*/
ods noproctitle;
ods graphics / imagemap=on;

proc means data=WORK.HIDATA chartype n nmiss vardef=df;
    var spq1 q2 q3 q5a iq5b q5c iq5d q5e iq5f q6 q7 q8 iq9a1 iq9a2 iq9a3 iq9a4
        iq9a5 iq9b1 iq9b2 iq9b3 iq9b4 iq9b5 slq9b q10 q11 q12_1 q12_2 q12_3 q12_4
        q12_5 q12_6 q12_7 q12_8 q12_9 q12_10 q12_11 q13 q14 exq15 iq17 exq18 q19a
        q19b q19c q19d q19e q19f q19g q19h q20spa q20spb q20spc q20spd q20spe q21a
        q21b q22 iq23 q24 iq25 iq26 q27 q28 q29 q30 q31 niq32 q33 niq34 q35 niq36
        niq37 q38 q39_1 q39_2 q39_3 q39_4 q39_5 q39_6 q39_7 q43 q44a q44b q44c q44d
        q44e q44f q44g q44h q44i q45a q45b q45c q45d q45e q45f q45g q45h q45i q46sp_1
        q46sp_2 q46sp_3 q46sp_4 q46sp_5 q46sp_6 q46sp_7 q46sp_8 q46sp_9 q46sp_10
        q46sp_11 q46sp_12 q46sp_13 q46sp_14 q46sp_15 q46sp_16 q46sp_17 q46sp_18
        q47sp_1 q47sp_2 q47sp_3 q47sp_4 q47sp_5 q47sp_6 q47sp_7 q47sp_8 q47sp_9 q48a
        q48b q48c q48d q49a q49b q49c q49d q49e q50 q52 q53 q54a q54b q55 q58 q58_2
        q59a q59b q63b sipaq bmi absi qevi qpmhp ac metrc_1 metrc_2 metrc_3 region
        urbrul dep key1 NS_SEC3 ageclass agecls2 agecls3 edu ctrybrth socldgt mainwgt
        bmiwgt;
run;

/*removal of vars with majority of missing values and derived variables: qevi qpmhp*/
data source.hidata_nomissing; set hidata
    (drop = iq5d iq5f iq9a1 iq9a2 iq9a3 iq9a4 iq9a5 iq9b1 iq9b2
        iq9b3 iq9b4 iq9b5 niq32 q11 q12_1 q12_10 q12_11 q12_2 q12_3
        q12_4 q12_5 q12_6 q12_7 q12_8 q12_9 q13 q58_2 q59b q8 slq9b
        qevi qpmhp );
run;

```

```

/*summary statistics of remaining data */
ods noproctitle;
ods graphics / imagemap=on;

proc means data=SOURCE.HIDATA_NOMISSING chartype mean std min max median
    vardef=df skewness kurtosis qmethod=os;
    var spq1 q2 q3 q5a iq5b q5c q5e q6 q7 q10 q14 exq15 iq17 exq18 q19a q19b q19c
    q19d q19e q19f q19g q19h q20spa q20spb q20spc q20spd q20spe q21a q21b q22
    iq23 q24 iq25 iq26 q27 q28 q29 q30 q31 q33 niq34 q35 niq36 niq37 q38 q39_1
    q39_2 q39_3 q39_4 q39_5 q39_6 q39_7 q43 q44a q44b q44c q44d q44e q44f q44g
    q44h q44i q45a q45b q45c q45d q45e q45f q45g q45h q45i q46sp_1 q46sp_2
    q46sp_3 q46sp_4 q46sp_5 q46sp_6 q46sp_7 q46sp_8 q46sp_9 q46sp_10 q46sp_11
    q46sp_12 q46sp_13 q46sp_14 q46sp_15 q46sp_16 q46sp_17 q46sp_18 q47sp_1
    q47sp_2 q47sp_3 q47sp_4 q47sp_5 q47sp_6 q47sp_7 q47sp_8 q47sp_9 q48a q48b
    q48c q48d q49a q49b q49c q49d q49e q50 q52 q53 q54a q54b q55 q58 q59a q63b
    sipaq bmi absi ac metrc_1 metrc_2 metrc_3 region urbrul dep key1
    NS_SEC3 ageclass agecls2 agecls3 edu ctrybrth socldgt mainwgt bmiwgt;
run;

/*export to csv to allow for loading into WEKA for visualisation of distribution and
feature selection*/
proc export data=source.hidata_nomissing dbms = csv
    outfile= '/folders/myfolders/hidata_nomissing.csv';
run;

/*SAS dataset with selected features only - selection made based on WEKA outputs*/
data source.hidata_selected; set source.hidata_nomissing
(keep = q45d q45a q45e q45b q45g q45i spq1 q46sp_7 q3 q2 q46sp_8 Sipaq q44c q54a q43
q46sp_9 q31 q24 q44b q58 niq37 q5e q46sp_16 q45h_class);
id = _n_;
run;

/*summary statistics of selected features */
ods noproctitle;
ods graphics / imagemap=on;

proc means data=source.hidata_selected chartype mean std min max median
    vardef=df skewness kurtosis qmethod=os;
    var q45d q45a q45e q45b q45g q45i spq1 q46sp_7 q3 q2 q46sp_8 Sipaq q44c q54a q43
q46sp_9 q31 q24 q44b q58 niq37 q5e q46sp_16 ;
run;

/*export to csv to allow for loading into WEKA for normalization and SMOTE */
proc export data=source.hidata_selected dbms = csv
    outfile= '/folders/myfolders/hidata_selected.csv';
run;

/*import of datasets: 'normalized' and 'normalized and balanced' in WEKA */
proc import datafile = '/folders/myfolders/hidata_selected_norm.csv'
    out=source.hidata_selected_norm dbms = csv;
run;

proc import datafile = '/folders/myfolders/hidata_selected_norm_balanced.csv'
    out=source.hidata_selected_norm_balanced dbms = csv;
run;

```

```

/*summary statistics of 'noramlized and balanced' dataset */
ods noproctitle;
ods graphics / imagemap=on;

proc means data=source.hidata_balanced chartype mean std min max median
    vardef=df skewness kurtosis qmethod=os;
    var q45d q45a q45e q45b q45g q45i spq1 q46sp_7 q3 q2 q46sp_8 Sipaq q44c
    q54a q43 q46sp_9 q31 q24 q44b q58 niq37 q5e q46sp_16;
run;

/*summary statistics 'noramlized' datasets */
ods noproctitle;
ods graphics / imagemap=on;

proc means data=source.hidata_balanced_norm chartype mean std min max median
    vardef=df skewness kurtosis qmethod=os;
    var q45d q45a q45e q45b q45g q45i spq1 q46sp_7 q3 q2 q46sp_8 Sipaq q44c q54a q43
    q46sp_9 q31 q24 q44b q58 niq37 q5e q46sp_16;
run;

/*target distribution visualisation*/

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=SOURCE.HIDATA_SELECTED_NORM;
    title height=14pt "Before SMOTE";
    vbar q45h_class / datalabel;
    yaxis grid;
run;

ods graphics / reset;
title;

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=SOURCE.HIDATA_SELECTED_NORM_BALANCED;
    title height=14pt "After SMOTE";
    vbar q45h_class / datalabel;
    yaxis grid;
run;

ods graphics / reset;

```

```

/*validation result join into one dataset to allow for t-test and visualisation*/
libname source '/home/d111240850/MscThesis';

proc sql;
    create table source.dt1_train_misc
    as select _fold_,
        sum(
            case
            when F_q45h_class =I_q45h_class then 1
            else 0
            end) as TP_count,
        count(F_q45h_class) as obs_count
    from source.dt1_train
    group by _fold_;

data source.dt1_train_misc; set source.dt1_train_misc;
Misc = 1 - (TP_count/obs_count);
run;

proc export data=source.dt1_train_misc dbms = csv
outfile= '/home/d111240850/MscThesis/dt1_train.csv';
run;

proc sql;
    create table source.dt2_train_misc
    as select _fold_,
        sum(
            case
            when F_q45h_class =I_q45h_class then 1
            else 0
            end) as TP_count,
        count(F_q45h_class) as obs_count
    from source.dt2_train
    group by _fold_;

data source.dt2_train_misc; set source.dt2_train_misc;
Misc = 1 - (TP_count/obs_count);
run;

proc export data=source.dt2_train_misc dbms = csv
outfile= '/home/d111240850/MscThesis/dt2_train.csv';
run;

proc sql;
    create table source.nn1_train_misc
    as select _fold_,
        sum(
            case
            when F_q45h_class =I_q45h_class then 1
            else 0
            end) as TP_count,
        count(F_q45h_class) as obs_count
    from source.nn1_train
    group by fold ;

```

```

data source.nn1_train_misc; set source.nn1_train_misc;
Misc = 1 - (TP_count/obs_count);
run;

proc export data=source.nn1_train_misc dbms = csv
outfile= '/home/dl11240850/MscThesis/nn1_train.csv';
run;

proc sql;
create table source.nn2_train_misc
as select _fold_,
sum(
case
when F_q45h_class =I_q45h_class then 1
else 0
end) as TP_count,
count(F_q45h_class) as obs_count
from source.nn2_train
group by _fold_;

data source.nn2_train_misc; set source.nn2_train_misc;
Misc = 1 - (TP_count/obs_count);
run;

proc export data=source.nn2_train_misc dbms = csv
outfile= '/home/dl11240850/MscThesis/nn2_train.csv';
run;

proc sql;
create table source.results
as select a._fold_,
a.Misc as Misc_dt1,
b.Misc as Misc_dt2,
c.Misc as Misc_nn1,
d.Misc as Misc_nn2
from source.dt1_train_misc a
left join source.dt2_train_misc b on a._fold_ = b._fold_
left join source.nn1_train_misc c on a._fold_ = c._fold_
left join source.nn2_train_misc d on a._fold_ = d._fold_
;
quit;

proc export data=source.results dbms = csv
outfile= '/home/dl11240850/MscThesis/results.csv';
run;

```