**Project Proposal**

Investigation Into Factors That Affect Wellbeing Using Machine Learning

Student ID: 10740709

Program: MSc Data Science and Business Analytics

Module: PROJ518 MSc Dissertation and Research Skills

Supervisor: Malgorzata Wojtys

Table of Contents

[1 Introduction 3](#_Toc99184360)

[2 Aims, Objectives and Methodology 4](#_Toc99184361)

[3 Relevant Literature 6](#_Toc99184362)

[5 Risk Assessment 7](#_Toc99184363)

[6 Ethics 8](#_Toc99184364)

[7 Impact 8](#_Toc99184365)

[8 Project Plan 9](#_Toc99184366)

[9 References 10](#_Toc99184367)

# 1 Introduction

Wellbeing exists as a multifaceted term which, at its core, seeks to define elements of intrinsic value to an individual. These elements combine to paint a picture of how well an individual can function and how they feel about their day-to-day life (Michaelson, Mahony and Schifferes, 2012).

Elements of wellbeing can be broadly categorised as either subjective (**SWB**), or objective (**OWB**). SWB is best understood as a person’s rating of their own life and ability to engage in activities that they value (Stone et al., 2013, p.5). This is a uniquely personal viewpoint, two people in a seemingly identical scenario could hold wildly differing opinions towards said scenario. OWB is harder to define given the multitude of dimensions that it covers. Voukelatou et al. (2020) suggest that OWB can be considered as the elements essential to driving wellbeing at a societal level. They go on to introduce the following as measurable dimensions of OWB: health, job opportunities, socioeconomic development, environment, safety and politics.

In 2010 the Office of National Statistics (**ONS**), began the ‘Measuring National well-being’ program in the UK. Everett (2015) explains the aim of this program is to develop a set of statistics to better equip people to understand and monitor wellbeing in the UK, this goes beyond previous, limited, measures of a nations progress such as GDP (gross domestic product). Everett continues to highlight the benefit of SWB measures and introduces 4 questions that have been added to ONS household surveys in order to further this program:

* Overall, how satisfied are you with your life nowadays?
* Overall, to what extent do you feel the things you do in your life are worthwhile?
* Overall, how happy did you feel yesterday?
* Overall, how anxious did you feel yesterday?

Answers to these questions are provided on a scale of 0 to 10, 0 is ‘not at all’ and 10 is ‘completely’. These questions will form the basis, or dependant variables, of the research posed in this project.

For this project, data collected by the ONS in the ‘national annual survey’ will be utilised to explore and identify factors that affect the SWB of the British population (Office for National Statistics, Social Survey Division, 2022). Contributory factors will be further analysed providing a view of how much each element adds to or detracts from measures of SWB.

Data consists of the 4 SWB questions alongside an additional 517 variables covering education, employment, health, ethnicity and more. Results are provided for 217,666 persons representing each local and unitary authority within the UK.

# 2 Aims, Objectives and Methodology

The overarching aim of this study will be to develop a predictive machine learning (ML) model alongside data visualisations displaying how various factors influence SWB. It will seek to simplify comprehension of the data, providing clear and understandable interpretations of identified relationships. Output will be of benefit in several areas such as personal decision making (e.g., will an individual be happier if they earn more? less anxious if they have more job opportunities?) and government policy making (e.g., does improved education lead to a happier populace?).

To achieve these aims the following initial objectives are proposed:

* Literature research
  + Identify and review literature relevant to the study. Including similar projects, wellbeing concepts, ML, data analysis methods and concepts
* Data cleansing
  + Variables deemed irrelevant to the study will be removed
  + Variables contain missing entries. Imputation techniques will be explored and implemented, where necessary, to provide a more complete data set
  + Variable values are numeric but relate to descriptive labels. These values will be translated to their descriptors
* Exploratory data analysis
  + Visualise variables exploring initial relationships to SWB measures and overall relevance to study
  + Feature selection. Identify the key variables for further exploration/removal
  + Explore how SWB measures relate to one another
  + Experiment with clustering of similar participants
  + Explore combining individual SWB variables into an overall measure

Initial objectives will take place simultaneously forming the bulk of the research project. Thorough early exploration and research will help contextualise the data within the field of wellbeing informing decision making for the development stages of the project.

These first steps are key to ensuring preparedness for the latter stages of the research.

The latter stages of the project will focus on production of ML models, data outputs and final write up of the project:

* ML experimentation
  + Trial ML methods with identified key variables and assess results alongside project aims
  + Decide upon ML methods for final models and outputs
* Design and execute ML methods
  + Build the final ML models. Execute, troubleshoot and refine them
* Design and produce outputs
  + Decide on desired visual outputs for ML and analysis representation
  + Build outputs with a priority on comprehension of identified relationships between variables and SWB

Clustering(grouping) of participants by certain meaningful variables, such as levels of income and education, will be explored. It will then be possible to identify how variables effect the SWB of these individual clusters of participants (Bruce, Bruce and Gedeck, 2020, p.283). This will not only make the data easier to comprehend, but it may also provide a clearer insight of specific wellbeing scenarios within the specifically defined clusters.

All code will be developed in R Studio and will utilise several relevant libraries such as caret, ggplot2 and dplyr. Additional libraries relevant to specific ML algorithms will also be utilised.

Considerable time has been set aside for data cleansing and exploratory data analysis as this is an extremely important step of any data project (Bruce, Bruce and Gedeck, 2020, p.46).

# 3 Relevant Literature

Prior to attempts to analyse the data, it is prudent to seek a better understanding of wellbeing, this will help contextualise findings as well as provide an early indication of variables that may be relevant from the data. Dodge et al. (2012) explores perspectives on the definition and dimensions of wellbeing from initial writings on the subject to the modern. It argues that previous research has focussed primarily on various dimensions of wellbeing rather than seeking to cement a clear definition of the term. One notable theory highlighted in the article is that of ‘equilibrium’ (p. 226-227) which then expands into ‘challenges’ (p. 228), these posit that challenges faced in life can have a varyingly detrimental effect on levels of SWB. The strength of this detrimental effect is determined by scale of the challenge in addition to the individuals ‘homeostatic defence’, this can be better understood as the elements that make up the average wellbeing balance of an individual. Where this balance is off, the defence is considered compromised, any challenges are likely to have a higher detrimental effect on an individual’s SWB. The article concludes with a new definition of wellbeing which is largely influenced by the theories discussed. Wellbeing relies on an individual’s resources (social, psychological, physical) being at a level adequate to deal with the challenges of life, where the demands of challenges outweigh the resources, wellbeing suffers as a result. This definition will be important to consider when assessing the data variables, if the variables represent departure from the normal balance of an individual (recent job loss, drop in income, etc) this is likely to correlate with lower levels of SWB. Whilst this article provides a good foundation to base an understanding of what wellbeing is and how elements may negatively affect SWB, it doesn’t explore in detail the elements that may increase SWB. It will be of benefit to investigate additional literature that focusses on elements that may increase an individual’s SWB in both the long and short term.

Alongside an understanding of what wellbeing is, findings such as the Easterlin Paradox provide valuable insight on the effect elements have on SWB. Easterlin et al. (2010) suggests that, over the long-term, ‘happiness does not increase as a country's income rises’. This also directly relates to individuals, past a certain base level of income, levels of happiness do not notably increase. This suggests it will not be as simple as looking at one variable alongside SWB measures and inferring a correlation. Consideration of long/short term effects will be needed as well as looking at multiple variables simultaneously.

Reviewing similar research projects is prudent to developing early ideas of suitable ML methods, it may also provide an early indication of elements from the data worth investigating. Makridis et al. (2021) seeks to investigate how various demographic, socio-economic and geographic characteristics contribute to the physical health and wellbeing of veterans. The study combines data from polls with census data to build a predictive model of wellbeing. Several algorithms are developed for each dataset: ordinary least squares linear regressor (OLS), XGBoost regressor, multi-layer perceptron regressor (MLP), support vector regressor (SVR), logistic regression (LR) classifier and a support vector classifier (SVC). Results of the study suggest financial anxiety results in stress that, in turn, negatively influences wellbeing of participants. It also highlights the effect education, age, race and gender have as important predictors of low levels of wellbeing. SWB classifiers developed in the study were relatively poor suggesting notable challenges in predicting overall wellbeing when compared to physical wellbeing.

In Naixin et al. (2019), a gradient boosting classifier is utilised to predict SWB in adolescents. 20 factors were identified from a series of psychological and physiological parameters as beneficial and used in the model resulting in a 90% correct prediction rate. These factors were narrowed down from a series of 298 with elastic net regression to avoid overfitting the model. This study focuses more on subjective measures such as feelings of depression and personality as predictors for SWB rather than objective measures such as education, gender, etc. In addition, SWB in this study is simply classified as ‘happiness’ which could be considered a narrow perspective of wellbeing. ML methods proposed by this paper do however show extremely positive results and are therefore well worth further consideration.

Kaur et al. (2019) proposes a predictive ensemble ML model for quality-of-life scoring. Initially, multiple ML methods were employed and evaluated including decision trees (DTs), neural network (NN), random forest (RF), SVR, cubist regression modelling, generalised linear model (GLM) and elastic net. 3 of the individually best performing models combine in the ensemble model boasting a 90% accurate prediction rate of OECD better life index scores (BLI). A combined model is demonstrated as performing better than the next best individual model, SVR. Kaur et al. aims to accurately predict wellbeing scores for a country rather than individuals, this poses different challenges to those that will likely arise during the research suggested in this proposal. Despite differences, ML methods and performance findings are relevant for consideration and an ensemble model should be explored.

# 5 Risk Assessment

|  |  |  |  |
| --- | --- | --- | --- |
| Risk | Likelihood | Impact | Mitigating Actions |
| Loss of project code | Low | Extreme | Project will be hosted on GitHub, stored on primary IT device and on the cloud (OneDrive) |
| Lack of available time due to other demands, illness etc | Low | High | Dedicated time will be allocated to the project each day. In the event of temporary unavailability, time will be made up later. Project plan will be monitored and updated accordingly. Ample time provided for each step. |
| Difficulty of task | Medium | High | Ample time provided for each step. Support available from peers and project supervisor. |
| Scope creep | Low | Low | Project plan and desired outcomes have been well defined. Minor changes will be manageable if deemed necessary. |
| Failure of IT equipment | Low | Low | Personal PC more than sufficient in reliability and power for task. Backup home device available if needed. Should performance be an issue, access to university devices can be sought. |

# 6 Ethics

This research will solely use data obtained online via the ONS (Office for National Statistics, Social Survey Division, 2022). Data is both anonymous and publicly available. For this reason, ethical approval will not be required.

# 7 Impact

Results obtained during this research carry potential for notable impact. Should models be developed that serve as a reliable predictor of subjective wellbeing, these could form the basis of suggestions for local government, NHS and other organisations.

Referencing the UN’s sustainable development goals, goal 3 relates to promoting healthy lives and wellbeing (United Nations, 2021). This research may serve to highlight key factors in improving and maintaining wellbeing in a post pandemic world.

Some key steps should be considered in order to achieve this potential impact:

1. Ensure models developed are robustly tested and scrutinised to a high standard
2. Proactively publish and share key findings with agencies such as the ONS, UK Health Security Agency (UKHSA) etc
3. Seek to formally present findings at relevant data science, health and wellbeing conferences/seminars

# 8 Project Plan

Chart

Description automatically generated with medium confidence

# 9 References

Bruce, P.C., Bruce, A. and Gedeck, P. (2020). Practical statistics for data scientists : 50+ essential concepts using R and Python. Sebastopol, Ca: O’reilly Media, Inc, p.46, 283.

Dodge, R., Daly, A., Huyton, J. and Sanders, L. (2012). The Challenge of Defining Wellbeing. International Journal of Wellbeing, [online] 2(3), pp.222–235. Available at: https://www.internationaljournalofwellbeing.org/index.php/ijow/article/view/89/238 [Accessed 20 Mar. 2022].

Easterlin, R.A., McVey, L.A., Switek, M., Sawangfa, O. and Zweig, J.S. (2010). The happiness-income paradox revisited. *Proceedings of the National Academy of Sciences*, [online] 107(52), pp.22463–22468. Available at: https://www.pnas.org/doi/10.1073/pnas.1015962107 [Accessed 20 Mar. 2022].

Everett, G. (2015). Measuring National Well-Being: A UK Perspective. Review of Income and Wealth, [online] 61(1), pp.34–42. Available at: http://www.roiw.org/2015/n1/05%20-%2012175.pdf [Accessed 14 Mar. 2022].

Kaur, M., Dhalaria, M., Sharma, P.K. and Park, J.H. (2019). Supervised Machine-Learning Predictive Analytics for National Quality of Life Scoring. *Applied Sciences*, [online] 9(8). Available at: https://www.mdpi.com/2076-3417/9/8/1613 [Accessed 20 Mar. 2022].

Makridis, C., Zhao, D., Bejan, C. (Adi) and Alterovitz, G. (2021). Leveraging Machine Learning to Characterize the Role of Socio-Economic Determinants on Physical Health and Well-Being Among Veterans. Computers in Biology and Medicine, [online] 133(104354). Available at: https://gpsych.bmj.com/content/32/5/e100096 [Accessed 20 Mar. 2022].

Michaelson, J., Mahony, S. and Schifferes, J. (2012). Measuring Well-being: A guide for practitioners. 1st ed. [online] New Economics Foundation, p.6. Available at: https://b.3cdn.net/nefoundation/8d92cf44e70b3d16e6\_rgm6bpd3i.pdf [Accessed 14 Mar. 2022].

Naixin, Z., Chuanxin, L., Lin, A., Zhixuan, C., Decheng, R., Fan, Y., Ruixue, Y., Lei, J., Yan, B., Zhenming, G., Gaini, M., Fei, X., Fengping, Y., Liping, Z., Gabriel, R., Yifeng, X., Lin, H., Bo, B., Guang, H. and Tao, Y. (2019). Prediction of Adolescent Subjective Well-Being: A Machine Learning Approach. Mathews Journal of Psychiatry & Mental Health, [online] 4(1). Available at: https://gpsych.bmj.com/content/32/5/e100096?utm\_campaign=gpsych&utm\_content=consumer&utm\_medium=cpc&utm\_source=trendmd&utm\_term=usage-042019 [Accessed 30 Aug. 2020].

Office for National Statistics, Social Survey Division, 2022. Annual Population Survey, October 2020 - September 2021, [data collection], UK Data Service, Accessed 12 March 2022. SN: 8886, DOI: 10.5255/UKDA-SN-8886-1 (http://doi.org/10.5255/UKDA-SN-8886-1)

Stone, A.A., Mackie, C.J., Policy-Relevant, I.A. and National Research Council (U.S.). Committee On National Statistics (2013). Subjective well-being : measuring happiness, suffering, and other dimensions of experience. [online] Washington, D.C.: The National Academies Press. Available at: https://www.multiplechronicconditions.org/assets/pdf/Depression%20Guidelines/Academies\_Press\_2013\_Subjective\_Well-Being\_Measuring\_Happiness\_Suffering\_and\_Other\_Dimensions\_of\_Experience.pdf [Accessed 21 Mar. 2022].

United Nations (2021). Goal 3 | Department of Economic and Social Affairs. [online] sdgs.un.org. Available at: https://sdgs.un.org/goals/goal3 [Accessed 14 Mar. 2022].

Voukelatou, V., Gabrielli, L., Miliou, I., Cresci, S., Sharma, R., Tesconi, M. and Pappalardo, L. (2020). Measuring objective and subjective well-being: dimensions and data sources. International Journal of Data Science and Analytics, [online] 11(279–309). Available at: https://link.springer.com/article/10.1007/s41060-020-00224-2 [Accessed 15 Mar. 2022].