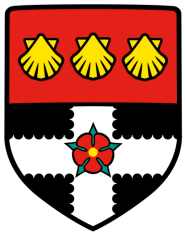

University of Reading

Department of Computer Science

Feature Efficient Prediction of Subjective Wellbeing in the UK Using Interpretable Machine Learning Models

Yiu Ting Wong

Supervisor: Dr. Ferran Espuny-Pujol

A report submitted in partial fulfilment of the requirements of
the University of Reading for the degree of
Master of Science in *Data Science and Advanced Computing*

September 11, 2025

Declaration

I, Yiu Ting Wong, of the Department of Computer Science, University of Reading, confirm that this is my own work and figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person's work, except where the works of others have been explicitly acknowledged, quoted, and referenced. I understand that if failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalised accordingly.

I give consent to a copy of my report being shared with future students as an exemplar.

I give consent for my work to be made available more widely to members of UoR and public with interest in teaching, learning and research.

Yiu Ting Wong

September 11, 2025

Abstract

Wellbeing is increasingly recognized as a core indicator of societal progress in the UK, yet official analyses such as the Office for National Statistics (ONS) regression based reports typically rely on linear models and seldom evaluate out of sample performance. This dissertation aims to (i) predict personal wellbeing accurately while (ii) identifying a minimal, interpretable set of predictors and (iii) explaining model behaviour. Using the UK Annual Population Survey (APS, 2021–2023; N = 341,465), wellbeing is defined as the mean of the ONS personal wellbeing items like life satisfaction, worthwhileness, and happiness, then binarized at 8 into "Low–High" vs "Very High", yielding near balanced classes. After cleaning and encoding, features were ranked by normalized mutual information and pruned for redundancy ($|r| \geq 0.90$), producing a compact seven feature set. We benchmarked logistic regression, ridge, linear SVM, k-nearest neighbours, random forest, XGBoost, and a multilayer perceptron (MLP); ROC–AUC was the primary metric. XGBoost performed best (AUC = 0.758; Accuracy/F1 $\approx$ 0.698), with MLP and Random Forest essentially tied (AUC = 0.757). SHAP analyses showed consistent drivers across model families: anxiety dominated, followed by activity limitations and partnership/employment indicators; income like variables played a smaller role. Results demonstrate that robust wellbeing classification is achievable with a small, interpretable feature set, offering an evidence base for targeted interventions and more efficient survey designs. We note limitations around subgroup fairness and generalization, and outline extensions to longitudinal data, richer targets, and fairness auditing.

Keywords: United Kingdom, wellbeing, machine learning, classification, SHAP.

Word count: 11950

GitLab link: https://github.com/maxemporor/MSc-project

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Ferran Espuny-Pujol, for their invaluable guidance, feedbacks, and constant encouragement throughout this project. Their expertise and support were essential in shaping both the direction and quality of this dissertation.

I am also thankful to the academic staff at the Data Science and Advanced Computing, University of Reading, for providing the resources, facilities, and learning environment that made this work possible.

Special thanks go to my friends and colleagues for their support, encouragement, and helpful discussions during the course of this project. I am equally grateful to my family, whose patience, understanding, and belief in me provided the motivation to complete this work.

Finally, I would like to acknowledge the UK Data Service and the Office for National Statistics (ONS) for granting access to the Annual Population Survey data, without which this research would not have been possible.

Contents

List of Figures

List of Tables

List of Abbreviations

APS	Annual Population Survey
AUC	Area Under the Curve
CDEI	Centre for Data Ethics and Innovation
CV	Cross-Validation
DPIA	Data Protection Impact Assessment
EDA	Exploratory Data Analysis
GDPR	General Data Protection Regulation
ICO	Information Commissioner's Office
KNN	K-Nearest Neighbours
ML	Machine Learning
MLP	Multi-Layer Perceptron
NMI	Normalised Mutual Information
ONS	Office for National Statistics
PMSW-21	Physical, Mental and Social Well-Being Scale (21-item)
R^2	Coefficient of Determination
ROC	Receiver Operating Characteristic
SF-36	36-Item Short Form Health Survey
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SMPCS	School of Mathematical, Physical and Computational Sciences
SVM	Support Vector Machine
WEMWBS	Warwick–Edinburgh Mental Well-Being Scale
WHO	World Health Organization
WHO-5	World Health Organization–5 Wellbeing Index
XAI	Explainable Artificial Intelligence
XGBoost	eXtreme Gradient Boosting

Chapter 1

Introduction

Motivation

This chapter sets the scene for the dissertation by motivating why subjective wellbeing matters for UK policy and research, and why a machine learning approach is timely. Wellbeing captures how people feel about their lives, not just clinical status. Recent shocks (the pandemic, cost of living pressures) have sharpened its policy relevance. Yet most applied studies still lean on linear models, rarely report out of sample accuracy, and often retain large, redundant predictor sets. Chapter 1 frames this gap and introduces the core proposition of the thesis: that feature efficient, interpretable machine learning models can provide more accurate and transparent insights into the determinants of wellbeing, using large scale survey data gathered in a period of substantial social and economic change.

Organization

The chapter proceeds as follows. Section 1.1 provides background and policy context for wellbeing and motivates data driven modelling. Section 1.2 states the problem and research questions. Section 1.3 sets out the aim and specific objectives. Section 1.4 outlines the solution approach: data, preprocessing, feature selection, modelling, evaluation, and interpretability. Section 1.5 summarizes the study's contributions and key achievements. Section 1.6 closes with the organization of the remainder of the report.

1.1 Background

One of the most meaningful indicators of a country's progress is the wellbeing of its citizens. Wellbeing extends beyond physical health to include mental, emotional, and social aspects of life. In recent years, researchers and policymakers have increasingly recognized its importance, valuing it as highly as traditional health outcomes (ISPOR, 2025). Unlike standard health measures, wellbeing reflects how people feel about their daily lives, their resilience in facing challenges, and their overall life satisfaction.

Despite its importance, wellbeing is a multifaceted and complex construct, shaped by interacting social, economic, and health related factors. This complexity makes it a challenging outcome to predict using traditional analytical approaches.

Recent social and economic disruptions in the UK including the COVID-19 pandemic, which reinforced existing inequalities across health, education, and society (Tapper et al., 2025), and the ongoing crisis of increasing of cost of living, which remains the leading public concern (Office for National Statistics, 2025) have further highlighted the importance of wellbeing as a national policy priority. The Office for National Statistics (ONS) has embedded wellbeing measurement at the core of its population monitoring strategy. However, most existing studies continue to rely on linear regression based approaches. This creates an opportunity for machine learning to generate richer, data driven insights into the determinants of wellbeing at scale.

At the same time, the rise of digital healthcare and advanced data science tools has created new opportunities to study wellbeing in ways that were not previously possible. Large scale survey datasets now capture rich information about people's lifestyles, habits, and health attributes. By applying machine learning techniques, researchers can model non-linear interactions, uncover complex patterns, and improve the prediction of subjective wellbeing.

1.2 Problem statement

Most existing wellbeing studies rely on descriptive statistics or linear regression. While these approaches provide valuable insights, they are limited in their ability to capture the non-linear and complex interactions that shape wellbeing, and they often involve large sets of predictors without adequately addressing redundancy (Oparina et al., 2022). The multidimensional nature of wellbeing demands more sophisticated modelling approaches that can uncover hidden patterns while also identifying a parsimonious set of strong predictors useful for policymakers.

In addition, current official monitoring frameworks such as those of the Office for National Statistics (ONS) have further limitations. Their analyses do not report formal model performance, making it unclear how accurate their findings are. Moreover, their methods are not fully reproducible because the underlying data and code are not publicly available, reducing transparency and preventing independent validation (Vassilev et al., 2019).

This gap motivates the present study, which investigates whether machine learning methods can provide more robust, interpretable, and feature efficient models of wellbeing in the UK. Using large scale survey data collected in the aftermath of the COVID-19 pandemic and during the ongoing cost of living crisis, the study explores whether machine learning can reveal new insights into the determinants of wellbeing.

Accordingly, this research is guided by two central questions:

1. What can machine learning methods reveal about the subjective wellbeing of UK residents?
2. Can these predictions, developed using novel approaches to feature efficient modelling and interpretability provide more accurate and transparent insights compared with traditional methods?

1.3 Aims and objectives

Aim: The project aims to design a predictive framework that balances accuracy with interpretability, ensuring models are not only effective but also practically meaningful for wellbeing research and policy. Specifically, it seeks to develop and evaluate interpretable machine learning models that can predict subjective wellbeing from large scale UK survey data, while identifying the minimal set of predictors required for strong performance.

Objectives

To achieve this aim, the study will:

1. **Prepare and preprocess the dataset** through cleaning, encoding, scaling, and balancing to ensure data quality and comparability.
2. **Train and compare multiple predictive models**, including logistic regression, random forests, support vector machines, gradient boosting, k-nearest neighbours, and neural networks.
3. **Apply systematic hyperparameter tuning** to optimize model performance across methods.
4. **Evaluate predictive performance** using accuracy, F1-score, and ROC-AUC to compare traditional and modern approaches.
5. **Apply explainable AI methods** (e.g., SHAP) to interpret model outputs, highlight the most influential predictors, and connect results to the wider wellbeing literature.

1.4 Solution approach

This project implements a structured machine learning (ML) pipeline on the UK Annual Population Survey (APS, 2021-2023) to predict subjective wellbeing (Office for National Statistics, 2024). The pipeline comprises 9 stages: (1) data collection, (2) cleaning, (3) outcome construction, (4) feature engineering and partitioning, (5) feature selection, (6) model training and hyperparameter tuning, (7) evaluation, (8) explainability with SHAP, and (9) final results. The overall design priorities feature efficient modelling, fair performance estimation, and transparent explanations of model behaviour. To support readability, Figure 1.1 summarizes the end to end workflow, from raw data to interpretable predictions.

1.4.1 Data and preprocessing

- **Data source and scope.** APS three year pooled dataset (Jan 2021-Dec 2023): 341,465 respondents, 459 attributes.
- **Cleaning.** Removed attributes with **more than 5% missing values**, to reduce bias from incomplete data.

- **Outcome construction.** Computed a composite wellbeing score as the mean of the ONS personal wellbeing items: life satisfaction (SATIS), worth (WORTH), and happiness (HAPPY), after addressing missing codes and taking row mean to create wellbeing score column. Converted this continuous score to a **binary outcome**: *Low-High* (< 8) vs. *Very High* (≥ 8), yielding a near balanced distribution (49% vs. 51%).

- **Feature engineering.** Recoded health condition fields into **binary indicators** (e.g., mental/anxiety, heart or circulation, autism) to avoid spurious ordinality and improve interpretability, retained native binary fields as 0/1.

- **Partitioning**. Performed stratified splits to preserve class balance: 70% train, 15% validation, 15% independent test. All feature selection and tuning used only train/validation and the test set was held out for final estimation.

- **Feature selection.** Ranked predictors using Normalized Mutual Information (NMI) with the target, applied correlation pruning ($|r| \geq 0.90$) to remove redundancy. A learning curve check with Random Forest showed performance plateau with a compact set, so we retained seven non redundant predictors: see Table 3.1 for full labels, scales and coding): **ANXIOUS** (yesterday's anxiety, 0–10), **LIMACT** (limitations due to a health problem), **LIMITK** (health condition limits the kind of work that can be done; binary), **MARDY6** (marital/partnership status), **MARCHK** (spouse/partner is a household member; binary), **REDACT** (duration that normal day-to-day activities have been reduced), and **DISCURR20** (current disability status).

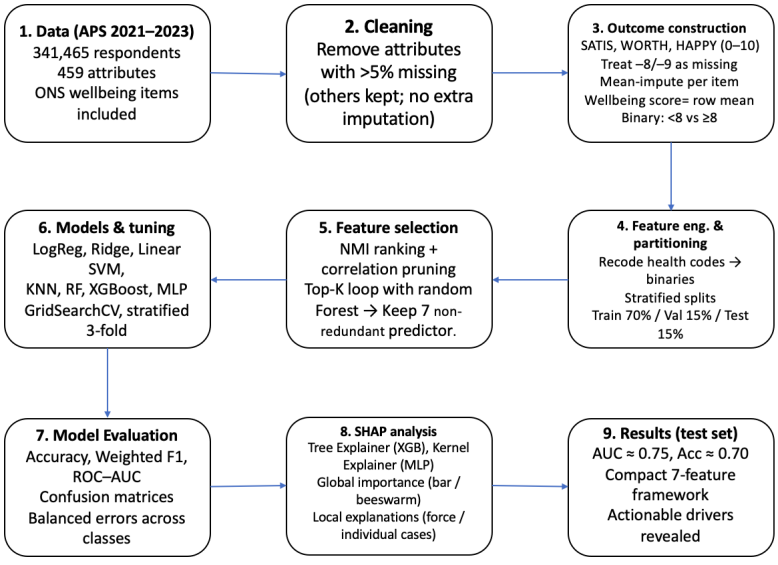

Figure 1.1: Overview of the solution pipeline

1.4.2 Modelling and interpretability

- **Algorithms.** Benchmarked a diverse family of models to balance interpretability and predictive strength: Logistic Regression, Ridge Classifier, Linear SVM, K-Nearest Neighbours (KNN), Random Forest, XGBoost, and Multi-Layer Perceptron (MLP).

- **Preprocessing by model.** Tree-based and linear models trained on raw (non-scaled) features, continuous variables were Z-scaled for MLP due to sensitivity to magnitude.
- **Hyperparameter optimization.** Used **GridSearchCV** with **stratified 3 fold cross validation** on the training set, selected configurations by **Accuracy**, **F1-score**, and **ROC-AUC** on the validation split; retrained the best model on **train+validation** before final testing.
- **Evaluation.** Reported **Accuracy**, **F1-score**, and **ROC-AUC,** inspected **confusion matrices** for error patterns and balance across classes.
- **Explainability.** We applied SHAP to the top-performing models, TreeExplainer for XGBoost and KernelExplainer for the MLP, using a stratified subsample for computational efficiency. We report global feature importance rankings and local bee swarm plots that show how individual cases are pushed toward each class. To aid interpretation, we also include targeted SHAP value plots for key variables, highlighting effect direction and magnitude. Together, these analyses link model attributions to established determinants of wellbeing and surface actionable insights.

1.5 Summary of contributions and achievements

This project makes four main contributions:

1. **Demonstrates the feasibility of wellbeing prediction using a compact set of predictors.** With only seven features retained after feature selection, the best performing model XGBoost achieved an AUC of 0.75, while an MLP reached 0.74. In contrast, simpler models such as Logistic Regression and KNN underperformed (AUC $\approx$ 0.68-0.70), highlighting the benefits of non-linear approaches without sacrificing interpretability.
2. **Establishes interpretability alongside predictive strength.** By applying SHAP explainability tools, the study shows that advanced models like XGBoost and MLP can produce transparent feature attributions. The findings confirm well established determinants such as anxiety and limiting long term illness or disability, but also elevate marital/partnership status (MARDY6) and recoded activity limitations (LIMACT, LIMITK) as consistently strong predictors.
3. **Identifies novel and policy relevant insights.** Unlike much of the existing literature, which emphasizes broad socio-economic correlates (e.g., income, education), this analysis highlights physical limitations and relationship status as equally or more influential than traditional economic variables. These findings suggest potential directions for targeted policy interventions for instance, wellbeing programmes focusing on managing chronic conditions or strengthening social support.
4. **Validates robustness across alternative modelling setups.** By testing alternative outcome splits (4 instead of 8), larger predictor sets (178 features), strict missing code policy, and alternative imputation strategies, the project demonstrates that the binary split at 8 combined with the 7 feature framework is the most stable, feature efficient, and interpretable configuration. This robustness enhances the external validity of the results and strengthens their practical utility for policymakers and healthcare providers.

Together, these contributions advance both the methodological toolkit of wellbeing research by showing that machine learning can offer feature efficient, reproducible models—and the applied evidence base, by highlighting a set of predictors that warrant further investigation for policy and practice.

1.6 Organization of the report

This report is organized into seven chapters:

- **Chapter 2** presents the literature review, including definitions of wellbeing, the UK context, determinants of wellbeing, and the role of machine learning.
- **Chapter 3** describes the methodology, including data collection, preprocessing, feature selection, algorithms, implementation, and ethical considerations.
- **Chapter 4** reports the results of model training, evaluation, and interpretability analysis.
- **Chapter 5** discusses the results in relation to wellbeing research, examines robustness, and considers limitations.
- **Chapter 6** concludes the study, summarizing achievements and outlining future work.
- **Chapter 7** provides a reflection on the learning process and skills gained throughout the project.

Chapter 2

Literature Review

Motivation

This chapter reviews the evidence base that underpins the study. It clarifies what is meant by subjective wellbeing, how it is commonly measured in the UK, and why recent social and economic shocks make an updated, policy relevant synthesis necessary. By contrasting the brevity and scalability of the ONS personal wellbeing questions with richer instruments (e.g., WEMWBS, WHO-5, SF-36), the chapter surfaces a key trade off between depth and coverage. It then examines what is known about determinants like health, relationships, employment, income, and housing and highlights where linear regression based approaches underperform, particularly for non-linear interactions and feature redundancy. Finally, it motivates the use of machine learning and explainable AI to combine predictive strength with transparency, while setting out the ethical and practical guardrails needed for responsible use.

Organization

Section 2.1 defines wellbeing and the UK measurement landscape, comparing the ONS framework to alternative instruments and justifying its use here. Section 2.2 situates wellbeing trends in the UK, focusing on the pandemic and cost of living pressures. Section 2.3 synthesizes evidence on social, psychological, economic, and health determinants. Section 2.4 reviews machine learning applications to wellbeing, comparative performance across algorithms, evaluation metrics, and the role of SHAP for interpretability. Section 2.5 offers a critique, identifying gaps this study addresses. Section 2.6 sets out ethical considerations across four angles, social impact, legal compliance, ethical principles, and risk management which is relevant to large scale survey modelling. Section 2.7 concludes with a brief summary to bridge to the methodology in Chapter 3.

2.1 Wellbeing research

Wellbeing is a broad concept that goes beyond the mere absence of illness. It includes physical, mental, emotional, and social dimensions, reflecting how people function and feel in their daily lives. Researchers' perspectives on wellbeing have been heavily influenced by the World Health Organization's well known definition of health as *"a state of complete physical, mental and social wellbeing, and not merely the absence of disease or infirmity."* (World

Health Organization, 1948). In practice, wellbeing can be interpreted through people's life satisfaction, resilience to hardships, sense of purpose, and the quality of their relationships. Because it reflects people's lives experiences, wellbeing is often considered one of the most meaningful indicators of how life is going in a country (Diener et al., 2018).

In the UK, the Office for National Statistics (ONS) has developed a set of personal wellbeing questions to monitor population wellbeing (Vassilev et al., 2019). These measures are used in large surveys such as the Annual Population Survey and ask individuals three core questions:

1. "Overall, how satisfied are you with your life nowadays?"
2. "Overall, to what extent do you feel the things you do in your life are worthwhile?"
3. "Overall, how happy did you feel yesterday?".

Responses are recorded on a scale from 0 to 10, and a composite wellbeing score is created by averaging the three answers. This score is then categorized into thresholds, with values of 0-4 labelled as *Low*, 5-6 as *Medium*, 7-8 as *High* and 9-10 as *Very High*.

The Office for National Statistics (ONS) wellbeing framework is widely used in the UK, but compared with more detailed measures, it has important limitations. Unlike tools such as the Warwick-Edinburgh Mental Well-Being Scale (WEMWBS), a 14-item scale focusing on positive aspects of psychological functioning (Tennant et al., 2007), or the WHO-5 Wellbeing Index, a brief yet extensively validated measure of general mental wellbeing (Topp et al., 2015), the ONS framework relies on only three short questions. Similarly, more holistic assessments such as the Physical, Mental and Social Well-Being Scale (PMSW-21), which combines physical health symptoms, emotional states, and social support (Supranowicz and Małgorzata, 2014), or the SF-36 Health Survey, which captures quality of life across eight domains including physical functioning and mental health (Ware and Sherbourne, 1992), provide a much richer view of wellbeing. In contrast, the brevity of the ONS questions means they cannot capture the deeper, multi-dimensional aspects of wellbeing, such as the causes of people's happiness or unhappiness, or subjective experiences like purpose, resilience, and community belonging.

However, the strength of the ONS framework lies in its **simplicity, scalability, and standardization**. It enables large scale comparisons across demographic and social groups at the national level and is highly practical for producing official statistics and monitoring wellbeing trends over time. These features make it particularly well suited for research that requires population level data.

For the purpose of this project, the ONS framework is the most suitable choice. Although it offers a less in depth understanding of wellbeing than more detailed instruments, its wide adoption, clarity, and consistency provide a robust foundation for predictive modelling with large scale UK survey data. This trade off between depth and scalability is acknowledged and taken into account when interpreting the findings of the present study.

2.2 Wellbeing in the UK

Between 2020 and 2023, personal wellbeing in the UK including life satisfaction, feelings of worthwhileness, and happiness experienced a significant decline, highlighting the ongoing effects of the COVID-19 pandemic (Office for National Statistics, 2022, 2023). These decreases reflect not just the immediate consequences of the health crisis but also prolonged societal anxiety and socioeconomic stress, particularly among vulnerable groups.

A further challenge to wellbeing has been the cost of living crisis, which began in late 2021. A study by King's College London found that 60% of respondents reported that rising living costs were harming their mental health, while 23% said they were experiencing sleep problems due to financial worries (King's College London, 2023). Additional research from Mind revealed that 48% of adults in England and Wales experienced negative mental health effects from the cost of living crisis, this increases to nearly 73% among those already living with a mental health condition (Mind, 2025). Kilfoyle (2023) show that financial hardship and debt correlate closely with poorer mental health, with projections estimating that energy cap increases could add significant numbers of people suffering from anxiety and depression.

2.3 Determinants of wellbeing

Wellbeing is shaped by a wide range of interrelated social, psychological, economic, and health factors. Social connections are among the most consistent predictors: higher levels of trust, supportive relationships, and civic engagement are strongly associated with greater subjective wellbeing (Helliwell and Putnam, 2004). At the individual level, psychological resources such as optimism and resilience play a key role in sustaining wellbeing, while good physical health and the absence of chronic illness remain fundamental determinants (Steptoe et al., 2015).

Economic conditions also exert significant influence. Income and employment provide material security, but their effects are complex: while higher income improves wellbeing up to the point of meeting basic needs, the positive effect diminishes beyond this threshold (Diener et al., 2018). Housing stability and the quality of work further shape both financial security and social identity, making them critical contextual drivers of wellbeing (Vassilev et al., 2019). Marital status and family circumstances have also been linked to life satisfaction, underscoring the importance of both relational and structural conditions.

Most of the evidence on these determinants comes from large scale surveys analyzed with regression based methods. These studies have consistently shown that health, relationships, and employment matter more for wellbeing than income alone (Powdthavee, 2008; Clark et al., 2018). However, regression models often explain only a small share of overall variation in wellbeing and struggle to capture non-linear interactions between determinants (Oparina et al., 2022). This limitation points to the value of applying machine learning approaches, which are better equipped to handle the multidimensional and interactive nature of wellbeing determinants.

2.4 Machine learning in wellbeing research

Machine learning (ML) approaches are now widely applied to wellbeing and health research, offering higher predictive accuracy and the ability to detect non-linear patterns compared with traditional models. Earlier studies often relied on simpler approaches such as logistic regression and decision trees, valued for their interpretability and ease of application. For example, Powdthavee (2008) used regression models to analyze wellbeing survey data, showing that factors like health and relationships outweighed income as predictors of life satisfaction. Similarly, Ferrer-i Carbonell and Frijters (2004) applied linear and panel regression techniques to wellbeing data, but found that their models explained only a small share of overall variance, highlighting the limits of such methods for prediction. These findings illustrate both the practical usefulness and the shortcomings of traditional statistical approaches when modelling wellbeing.

With advances in computational methods, more sophisticated ML models are increasingly applied. Random forests, for example, improve predictive power by combining multiple decision trees and are effective at handling large sets of wellbeing predictors (Abdul Rahman et al., 2023). Ensemble boosting methods such as XGBoost have demonstrated state of the art performance in survey-based wellbeing prediction, excelling in capturing complex feature interactions, handling class imbalance, and reducing overfitting through regularization (Oparina et al., 2022). Neural networks including multilayer perceptrons have been applied to wellbeing and mental health prediction with strong performance when diverse behavioural, physiological, demographic, and psychosocial features are included (Liu, 2024). Other advanced methods such as support vector machines (SVMs) and k-nearest neighbours (KNNs) have been used in smaller wellbeing-related studies, but they are often less scalable to high dimensional datasets (Islam et al., 2018).

Beyond applying individual models, some studies systematically compare multiple approaches to evaluate their relative performance in wellbeing prediction. For instance, Oparina et al. (2022) compared linear regression, random forests, and gradient boosting methods for predicting life satisfaction, and found that machine learning models substantially outperformed regression in terms of predictive accuracy and explained variance. Similarly, Abdul Rahman et al. (2023) tested a range of algorithms including logistic regression, k-nearest neighbours, naïve Bayes, neural networks, random forests, and boosting methods to predict mental wellbeing among university students, with ensemble approaches such as Random Forest and Adaptive Boosting achieving the strongest performance. These comparative studies highlight the value of benchmarking multiple models rather than relying on a single approach, as they reveal differences in predictive power, scalability, and interpretability.

In terms of evaluation, most studies report standard classification and regression metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) (Abdul Rahman et al., 2023; Oparina et al., 2022; Islam et al., 2018). Some also include variance explained (R^2) to assess overall fit. This diversity of metrics reflects the multidimensional goals of wellbeing research—not only to maximize predictive performance, but also to ensure that models capture meaningful patterns that can inform policy and practice.

A key development has been the rise of explainable AI (XAI) tools such as SHAP (SHAPley Additive Explanations), which enable interpretation of complex models like XGBoost and MLPs by identifying the relative importance of wellbeing predictors (Lundberg and Lee, 2017*a*; Lundberg et al., 2020). By providing both global feature rankings and local explanations, SHAP

has helped bridge the gap between accuracy and interpretability. Despite challenges such as algorithmic bias and trade offs between complexity and transparency, the growing evidence suggests that tree-based ensembles and neural networks, when combined with interpretability tools, represent some of the most effective modelling strategies for wellbeing research to date.

2.5 Critique of the review

The existing literature has established valuable measurement frameworks and highlighted consistent determinants of wellbeing, yet important gaps remain. National monitoring tools such as the ONS measures enable large scale tracking but lack the depth of richer instruments, which themselves are impractical for population wide use. Contextual analyses document declines in wellbeing during the pandemic and cost of living crisis, but these studies remain largely descriptive and fall short of predictive modelling. Traditional linear regression approaches further limit progress by assuming linear and independent effects, making it difficult to capture the complex interactions that shapes wellbeing. Although machine learning has begun to address these challenges, much of the work to date has relied on isolated models or emphasized predictive accuracy at the expense of interpretability. This underscores the need for approaches that combine robust predictive performance with transparency, ensuring results are both reliable and actionable for policy and practice.

Taken together, these gaps also highlight the timeliness of developing predictive approaches to wellbeing in the UK. The post pandemic recovery and the ongoing cost of living crisis have placed new pressures on resilience and life satisfaction, yet most national evidence remains descriptive. By applying machine learning to recent APS data (2021-2023), this project addresses these shortcomings directly, offering an up to date and policy relevant perspective on the factors associated with very high wellbeing.

2.6 Ethical Considerations

2.6.1 Social Impact

Data science based wellbeing prediction has the potential to create positive impacts for both individuals and society. For individuals, it can promote healthier behaviours by raising awareness of health and lifestyle factors that may influence wellbeing. On a wider scale, such predictive insights could guide communities and policymakers in developing strategies aimed at improving population wellbeing across the UK (HM Treasury, 2021).

By translating individual level data into meaningful insights, the project aligns with broader public health goals of resilience and quality of life. However, these benefits come with important cautions. Predictive models could be misused if results are taken out of context for example, in workplace monitoring, insurance profiling, or other forms of discrimination (Centre for Data Ethics and Innovation, 2020). There is also a risk of stigmatizing groups identified as having "low wellbeing."

2.6.2 Legal Compliance

This project is based on anonymized secondary data and carried out strictly within an academic setting. Even so, the handling of personal and health related information requires awareness of legal frameworks such as the General Data Protection Regulation (GDPR) (Information Commissioner's Office, 2023). These regulations ensure privacy, confidentiality, and robust data governance. While the dataset used here avoids direct identifiers, any future real world deployment of similar models would demand strict GDPR compliance, clear communication of dataset use, and careful respect for intellectual property rights surrounding both data and algorithms.

In practice, any future deployment of such models would require a lawful basis for processing (e.g., public interest or research), Data Protection Impact Assessments (DPIAs), and data minimization principles, ensuring that only necessary information is used. Moreover, datasets obtained through the UK Data Service are subject to strict licensing and intellectual property conditions, meaning their use outside academia would require additional permissions (UK Data Service, 2024). These measures together highlight the importance of strong legal compliance and transparent communication when scaling such predictive approaches.

2.6.3 Ethical Considerations

From an ethical perspective, the main concerns relate to privacy, fairness, and transparency. Although the APS dataset is anonymized, respondents did not explicitly consent to their information being repurposed for secondary research on wellbeing prediction, raising important questions about data use and participant expectations (Information Commissioner's Office, 2023).

Fairness is another key issue. Predictive analyses can inadvertently embed existing social and demographic biases, leading to underrepresentation of minority groups or systematic misclassification (Centre for Data Ethics and Innovation, 2020). Such outcomes risk reinforcing inequalities rather than alleviating them.

Transparency is also essential. Researchers must clearly report how data was processed, which variables were included, and the limitations of the findings. Overstating accuracy or drawing causal claims from observational survey data could mislead policymakers or the public (Collins et al., 2015).

To address these ethical challenges, this project emphasizes robust documentation, acknowledgement of dataset limitations, and cautious interpretation of findings. In applied settings, responsible use would further require independent oversight, clear communication with stakeholders, and equitable consideration of all groups affected by such analyses.

2.6.4 Risk Management

This project faced several technical and methodological risks that required careful management. A major challenge was dataset imbalance, which could bias predictions toward majority classes. To reduce this risk, the data was split into subsets in a way that preserved balanced distributions, rather than relying on synthetic oversampling methods such as SMOTE (Chawla

et al., 2002).

Model overfitting was another concern, addressed through cross validation, regularization, and systematic hyperparameter tuning (Varoquaux, 2018). Missing data was handled during preprocessing with appropriate exclusions and imputations, improving data quality.

Beyond these, other risks were also acknowledged. Concept drift poses a long term threat, as wellbeing predictors and distributions may change in the future, limiting the model's stability (Gama et al., 2014). Generalizability is also a challenge, as results derived from UK survey data may not transfer directly to other populations. Computational complexity, particularly in running SHAP for large models, introduced scalability risks. Finally, interpretability risks arise when highly complex models reduce transparency and public trust.

Together, these risks highlight the importance of cautious interpretation. While steps such as balanced sampling, feature selection, and explainable AI mitigated some risks, robustness and real-world reliability will ultimately require continuous monitoring and review.

2.7 Summary

In summary, existing literature confirms the importance of wellbeing as a multidimensional concept influenced by social, psychological, health, and economic factors. It also highlights the limitations of current measurement tools and analytic methods, especially in the UK context of recent social and economic crises. Machine learning offers a promising way forward, but its use has not yet been fully exploited in wellbeing research, particularly with respect to feature efficiency and interpretability. This project addresses these gaps by applying a comparative machine learning framework to large scale UK survey data, combining predictive accuracy with explainable outputs to provide insights that are relevant for both researchers and policymakers.

Chapter 3

Methodology

Motivation

This chapter sets out the methodological backbone of the study so that results in Chapter 4 are credible, reproducible, and easy to interpret. It explains how we transform the APS 2021–2023 data (Study 9291) into an analysis ready dataset, how we define the binary wellbeing outcome and how we reduce a high dimensional predictor space to a compact, non redundant set without leaking information from the test split. The approach emphasizes feature efficiency (NMI ranking + correlation pruning and a top-K check), fair performance estimation (stratified splits, held out test), and transparency (clear preprocessing steps, documented tuning, and SHAP based interpretation). Together, these design choices allow us to benchmark linear, tree-based, and neural models on equal footing and to connect predictive signals back to theory. For an end-to-end overview, see Figure 1.1.

Organization

Section 3.1 covers data preparation: survey scope and variables 3.1.1, outcome construction and binarizing 3.1.2, missingness and filtering 3.1.3, feature engineering and two stage selection with the top-K decision 3.1.4, and stratified partitioning 3.1.5. Section 3.2 details model development: software environment 3.2.1; algorithms and preprocessing policies 3.2.2; hyperparameter search via GridSearchCV with stratified 3-fold CV 3.2.3; evaluation metrics and selection criteria with ROC–AUC as primary 3.2.4; explainability via SHAP for tree based and MLP models 3.2.5; and the experimental protocol to ensure reproducibility and no peeking 3.2.6. Section 3.3 presents the full implementation. Section 3.4 addresses ethical and legal considerations (privacy, fairness, GDPR/licensing). Section 3.5 summarizes and links forward to the empirical results in Chapter 4.

3.1 Data preparation

3.1.1 Survey characteristics and variables

The dataset is the Annual Population Survey (APS) Three-Year Pooled Dataset (Study 9291), January 2021- December 2023 (Office for National Statistics, 2024): 341,465 respondents and 459 attributes. APS is a large scale household survey that captures socio-demographic,

economic, and health related indicators and includes the ONS personal wellbeing questions used for national monitoring.

Table 3.1: Key variables used in this study

Code	Label	Scale	Role in study
SATIS	Life satisfaction	0–10 (integer)	Outcome component
WORTH	Worthwhile	0–10 (integer)	Outcome component
HAPPY	Happiness yesterday	0–10 (integer)	Outcome component
ANXIOUS	Anxious yesterday	0–10 (integer)	Candidate predictor
LIMACT	Limitation due to health problem	Categorical	Candidate predictor
LIMITK	Limiting health problem affects kind of work can do	Binary	Candidate predictor
MARDY6	Married/Co-habiting/Civil Partners	Categorical	Candidate predictor
MARCHK	Whether spouse is household member	Binary	Candidate predictor
REDACT	How long ability to carry out normal day-to-day activities has been reduced	Categorical	Candidate predictor
DISCURR20	Current disability	Categorical	Candidate predictor

3.1.2 Outcome definition

The primary outcome is subjective wellbeing, measured using three APS items: life satisfaction (SATIS), sense of worth (WORTH), and happiness (HAPPY) (Table 3.1), each recorded on a $0-10$ scale. First, rows where all three items were missing were excluded. Next, APS special codes -8 ("No answer") and -9 ("Does not apply") were treated as missing (NaN). For each of the three items, missing values were replaced with the column mean to ensure comparability across respondents. A new variable, **Wellbeing**, was then created as the row wise average of SATIS, WORTH, and HAPPY. Finally, this continuous score was converted into a binary outcome, **Wellbeing_category**, with values below 8 labelled as *Low-High* and values of 8 or above labelled as *Very High*. This threshold yielded an approximately balanced class distribution, making it suitable for binary classification.

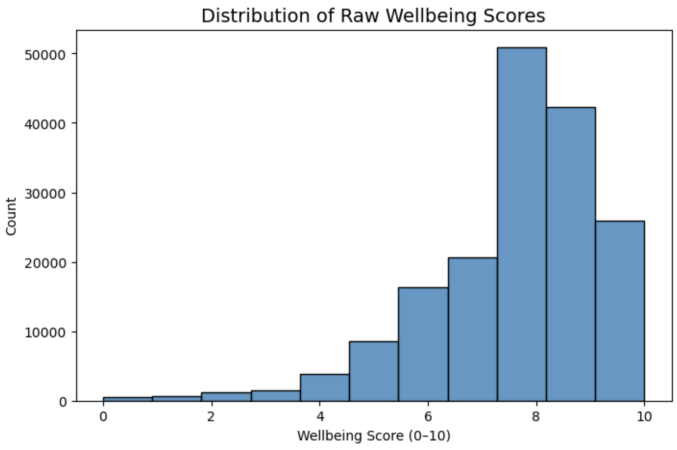

Figure 3.1: Distribution of Raw Wellbeing Scores

Figure 3.1 shows the distribution of the raw wellbeing score across respondents, while Figure 3.2 illustrates the resulting binary categories (*Low-High* vs. *Very High*). These figures demonstrate both the skew in the raw scale and the near balance of the derived binary outcome, which is important for ensuring a fair classification task.

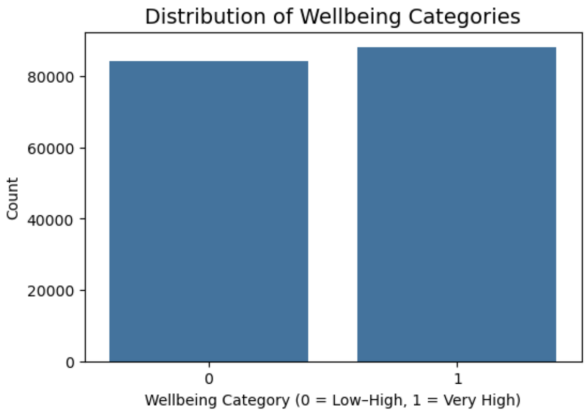

Figure 3.2: Distribution of Wellbeing Categories

3.1.3 Missing data and attribute filtering

An initial missing value analysis revealed that three variables: **GRSSWK** (Gross weekly pay in main job), **NETWK** (Net weekly pay in main job), and **RELIG11** (Religion, GB level) had more than 5% missing observations and were therefore excluded to avoid bias from highly incomplete attributes. For the remaining predictors, missingness did not exceed approximately 1.6%, so they were retained without additional imputation. The only imputation performed was the column mean imputation described above for SATIS, WORTH, and HAPPY to construct the composite outcome.

3.1.4 Feature engineering and selection

Health condition variables (HEALPB2001-HEALPB2010), which originally encoded conditions as numeric codes, were recoded into 18 binary indicators (e.g., *HEALTH_MentalAnxiety*, *HEALTH_HeartOrCirculation*, *HEALTH_Autism*). Binary encoding was chosen over ordinal representations to reduce noise, avoid false ordering, and improve interpretability. Redundant attributes such as identifiers (IDREF) and the original wellbeing items (SATIS, WORTH, HAPPY, Wellbeing) were removed. After preprocessing, the cleaned dataset contained **172,333 respondents** and **471 attributes**.

For feature selection, we adopted a two stage filter approach:

1. Normalized Mutual Information (NMI) was computed between each attribute Xj and the binary target Y to rank predictors. NMI rescales mutual information to the range [0,1], which makes features comparable (Vinh et al., 2010; Pedregosa et al., 2011).
2. Correlation pruning was applied among the ranked features, removing attributes with an absolute Pearson correlation $r >= 0.90$ (Kuhn, 2008).

Determining **K** (number of features). We then performed a top-K curve using a fixed classifier (Random Forest) to choose K at the point of performance plateau (performance defined in Section 3.2.4). We fixed K = 7 and carried forward the top seven non-redundant predictors (introduced in Table 3.1).

3.1.5 Partitioning

We used stratified splits to preserve class balance: 70% train, 15% validation, 15% independent test. All feature selection, model development, and tuning used only the train/validation data, the test set was held out once for unbiased final estimation.

3.2 Model development

3.2.1 Software and environment

All experiments were run in Python 3.12.4 (Anaconda) with JupyterLab 4.0.11. Key libraries: pandas, numpy (data), scikit-learn (preprocessing, CV/tuning), xgboost, SHAP, matplotlib (Pedregosa et al., 2011; Chen and Guestrin, 2016; Lundberg and Lee, 2017*b*).

3.2.2 Algorithms and preprocessing

We benchmarked a diverse family of models to balance interpretability and predictive capacity:

- Linear: Logistic Regression, Ridge Classifier, Linear SVM (Cortes and Vapnik, 1995)
- Distance-based: K-Nearest Neighbours (Cover and Hart, 1967)
- Ensembles: Random Forest (Breiman, 2001), XGBoost (Chen and Guestrin, 2016)
- Neural: Multi-Layer Perceptron (MLP) (Goodfellow et al., 2016)

Scaling policy. Tree-based and linear models were trained on raw features (scale-invariant or robust in practice), continuous variables were Z-scored for MLP due to sensitivity to feature magnitude.

3.2.3 Hyperparameter search

We used GridSearchCV with stratified 3-fold CV on the training set (Kohavi, 1995; Pedregosa et al., 2011). For each model, we selected the configuration that maximized the primary model selection criterion on the validation split (see section 4.3), then retrained on train+validation before the single test evaluation. For the detailed implementation of the grid-search procedure, see Section 3.3.8.

3.2.4 Evaluation metrics and selection criteria

We report Accuracy, weighted F1-score, and ROC-AUC:

- Accuracy is intuitive but can be misleading under even mild imbalance.

- Weighted F1 balances precision/recall across classes and penalizes asymmetric errors: useful when both false positives and false negatives matter (Sokolova and Lapalme, 2009).
- ROC-AUC is threshold independent and measures ranking/separability (Fawcett, 2006), Its used to compare modelling families robustly.

Model selection / K selection. We used ROC-AUC on the validation split as the primary selection metric for (1) choosing K in the top-K feature curve and (2) choosing tuned models within each family. Accuracy and weighted F1 were used as secondary checks for calibration and error balance.

3.2.5 Explainability methods

To interpret model behaviour, we applied SHAP (SHAPley Additive Explanations):

- TreeExplainer for tree-based models (e.g., XGBoost),
- KernelExplainer for MLP (with subsampling for tractability) (Lundberg and Lee, 2017*b*).

We produced global summaries (mean |SHAP| rankings) and local plots (e.g., bee swarm and Key feature SHAP plots) to characterize how feature values influence predictions. All empirical findings (e.g., which features dominate) are reported in Chapter 4.

3.2.6 Experimental protocol and reproducibility

- No peeking: Feature selection, tuning, and all interim evaluations were confined to train/validation data. The test set was untouched until the end.
- Reproducibility: Random seeds fixed for splits/tuning, software/library versions documented in 3.2.1, codebase archived in GitHub.

3.3 Implementation

3.3.1 Environment and tooling

All analyses were conducted in `Python`. The primary libraries were: `pandas` and `numpy` for data handling, `scikit-learn` for modelling and evaluation, `xgboost` for gradient-boosted trees, `matplotlib` for figures, and `shap` for post-hoc explainability. Reproducibility was ensured by fixing `random_state=42` across sampling and model training. Figures were exported automatically to a local `figures/` directory.

3.3.2 Data ingestion

The working dataset was the three-year APS extract (January 2021–December 2023), loaded from a tab-delimited file:

- **File:** `aps_3yr_jan21dec23_eul_withoutsmoking.tab`
- **Loader:** `pandas.read_csv(delimiter='\t', low_memory=False)`
- **Dtypes:** `Int64` for integer-coded variables (e.g. CLAIMS14) and `string` for text fields (e.g. CombinedAuthorities).

3.3.3 Target construction

Following ONS guidance, the three life-evaluation items `SATIS`, `WORTH`, and `HAPPY` (0–10) were processed as follows:

1. **Row exclusion:** remove rows where all three items were invalid (≤ -8).
2. **Item cleaning:** replace $-8/-9$ with the item mean (computed excluding these codes).
3. **Composite:** compute `Wellbeing` as the row-wise mean of SATIS, WORTH, and HAPPY.
4. **Binary label:** define `Wellbeing_category` as 1 if `Wellbeing` ≥ 8 (*Very High*), else 0 (*Low–High*).

3.3.4 Feature engineering

Health condition indicators

APS health problem indicators (`HEALPB2001`–`HEALPB2010`) were recoded into binary `HEALTH_*` flags (e.g. breathing problems, mental illness, progressive illness). Each flag equals 1 if the relevant code was present for that respondent.

Column removal

Variables removed prior to encoding:

- Outcome-related: `SATIS`, `WORTH`, `HAPPY`, `Wellbeing`
- Administrative: `IDREF`
- High missingness or out-of-scope: `GRSSWK`, `NETWK`, `RELIG11`

The resulting working table is denoted `df_filtered`.

Categorical encoding

All categorical variables in `df_filtered` were label-encoded, with missing entries temporarily filled with a sentinel "missing". Binary and numeric fields were retained unchanged. The encoded dataset is denoted `df_encoded`.

3.3.5 Train/validation/test partition

A stratified split was performed to preserve class prevalence:

- Train: 70%
- Validation: 15%
- Test: 15%

Splits were created with `train_test_split(..., stratify=Wellbeing_category, random_state=42)`.

3.3.6 Feature selection

Mutual information ranking

Normalised mutual information (NMI) was computed between each candidate predictor and the target. Features were ranked in descending order, with the top 20 visualized (Fig 4.1).

Redundancy pruning

To reduce multicollinearity, correlation pruning was applied: for $|r| \geq 0.90$, the lower-ranked feature was discarded. The procedure was logged and illustrated in a correlation heatmap (Fig 4.2).

Top-K curve

A diagnostic experiment evaluated Random Forest performance as features were added in NMI order (Fig 4.3).

3.3.7 Model exploration

Candidate models

Using the selected features, the following models were explored:

- Logistic Regression, Ridge Classifier

- Linear SVM, K-Nearest Neighbours
- Random Forest, XGBoost
- Multi-Layer Perceptron (MLP)

Scaling (`StandardScaler`) was applied only to the MLP. Validation metrics included Accuracy, weighted F1, and ROC–AUC. Per-model confusion matrices were produced.

3.3.8 Hyperparameter tuning and evaluation

Grid search cross-validation (3-fold, stratified) was applied separately for each model, optimising hyperparameters such as `n_estimators`, `max_depth`, `C`, `alpha`, and `hidden_layer_sizes`.

Hyperparameter grids. For transparency, the exact grids searched were:

- **Random Forest:** `n_estimators` $\in \{100, 200\}$; `max_depth` $\in \{\text{None}, 10, 20\}$; `min_samples_split` $\in \{2, 5\}$; `min_samples_leaf` $\in \{1, 2\}$ $\rightarrow$ **24** combinations.
- **XGBoost:** `n_estimators` $\in \{100, 200\}$; `max_depth` $\in \{3, 5, 7\}$; `learning_rate` $\in \{0.01, 0.05, 0.1\}$; `subsample` $\in \{0.8, 1.0\}$; `colsample_bytree` $\in \{0.8, 1.0\}$ $\rightarrow$ **72** combinations.
- **Ridge Classifier:** `alpha` $\in \{0.1, 1.0, 10.0\}$; `solver` $\in \{\text{auto}, \text{saga}\}$ $\rightarrow$ **6** combinations.
- **Logistic Regression:** `C` $\in \{0.1, 1.0, 10.0\}$; `penalty` $= \text{l2}$; `solver` $\in \{\text{lbfgs}, \text{saga}\}$ $\rightarrow$ **6** combinations.
- **Linear SVM (`LinearSVC`):** `C` $\in \{0.1, 1.0, 10.0\}$; `max_iter` $\in \{5000, 10000\}$ $\rightarrow$ **6** combinations.
- **KNN:** `n_neighbors` $\in \{3, 5, 7\}$; `weights` $\in \{\text{uniform}, \text{distance}\}$; `p` $\in \{1, 2\}$ $\rightarrow$ **12** combinations.
- **MLP:** `hidden_layer_sizes` $\in \{(64, 32), (128, 64)\}$; `activation` $\in \{\text{relu}, \text{tanh}\}$; `solver` $= \text{adam}$; `learning_rate_init` $\in \{0.001, 0.01\}$; `max_iter` $\in \{200, 400\}$ $\rightarrow$ **16** combinations.

Search budget. Across all models the grid spanned 142 hyperparameter combinations in total

The best estimator per model was refitted on Train+Validation and evaluated once on the Test set. Metrics were reported in descending ROC–AUC order (see Table 4.2). Normalised confusion matrices for all models are presented (Figure 4.4).

3.3.9 Explainability

SHAP values were used to interpret the top models:

- **XGBoost:** `shap.TreeExplainer` on the test set
- **MLP:** `shap.KernelExplainer` on a 100-row sample of scaled features

Global importance plots (mean absolute SHAP values, beeswarm) were generated. A custom routine aggregated SHAP contributions by raw survey response value, showing whether each pushed towards "Very High" or "Low–High" wellbeing. Figures are titled with the full survey question text for clarity (Figure 4.5, 4.6, 4.7).

3.3.10 Reproducibility and outputs

All splits and models were deterministic (`random_state=42`). Figures were exported as `.png` and `.pdf` into `figures/`, including:

- NMI bar chart (Figure 4.1)
- Correlation heatmap (Figure 4.2)
- Top-K feature curve (Figure 4.3)
- Per-model confusion matrices (Figure 4.4)
- SHAP plots(Figure 4.5, 4.6, 4.7)

Trained estimators can be serialised for replication.

3.4 Ethical and Legal Considerations

This project was conducted entirely with anonymized secondary data from the UK Data Service, ensuring that no direct identifiers were available at any stage of the analysis. Ethical risks related to privacy and fairness were mitigated by following data minimization principles, using only variables relevant to the research questions, and documenting all preprocessing and modelling decisions transparently. To address legal compliance, the study adhered to the licensing terms of the UK Data Service and observed principles of GDPR such as confidentiality and restricted use of sensitive information. Since this research is academic and exploratory, some broader ethical challenges (e.g., fairness in deployment, concept drift, and participant consent for secondary use) are acknowledged but not directly addressed within the project; these are discussed further as limitations in Chapter 5.

3.5 Summary

This chapter outlined the methodological framework and impletations used to investigate wellbeing prediction. The Annual Population Survey (2021-2023) was introduced, with key variables defined and the outcome constructed from ONS wellbeing measures. Data preparation included exclusion of highly incomplete attributes, column mean imputation for outcome

items, and binary encoding of health conditions. Feature selection was carried out using Normalized Mutual Information and correlation pruning, resulting in a compact, non redundant predictor set. The dataset was then partitioned into training, validation, and test subsets using stratified sampling to preserve class balance.

A range of modelling approaches was benchmarked, spanning interpretable linear models, distance-based methods, ensemble tree learners, and neural networks. Hyperparameter tuning was performed systematically with cross-validation, and models were evaluated using multiple complementary metrics (Accuracy, weighted F1, ROC-AUC) to ensure robustness. SHAP explainability was incorporated to support transparency across both tree-based and neural models. Throughout, strict experimental protocols were followed to prevent data leakage and ensure reproducibility.

Finally, this chapter addressed ethical and legal considerations. The project was conducted using anonymized secondary data in compliance with UK Data Service licensing and GDPR principles, with attention to privacy, fairness, and transparency. Broader challenges such as potential biases, concept drift, and implications for deployment are acknowledged as limitations to be revisited in Chapter 5.

Together, these methodological steps establish a rigorous and responsible foundation for the empirical results presented in Chapter 4.

Chapter 4

Results

Motivation This chapter presents the empirical results of the study and shows that a compact, theory aligned feature set can predict subjective wellbeing with reliable out of sample performance. We first verify that the target is well behaved and that redundant predictors are pruned, then demonstrate via model comparison and error analyses that non-linear methods (XGBoost, MLP, Random Forest) provide the best discrimination while maintaining balanced errors across classes. Finally, we use SHAP to connect predictive signals, especially anxiety and activity limitations to established determinants, ensuring findings are interpretable and decision relevant.

Organization Section 4.1 reports exploratory patterns and correlation based redundancy pruning. Section 4.2 details feature selection: NMI ranking, the correlation heat map, the top-K curve, and the resulting seven-feature specification. Section 4.3 shows the best results after model tunning. Section 4.4 compares models on validation and test set. Section 4.4.1 summarizes confusion matrices in a multi panel figure to assess error balance across classes. Section 4.5 provides interpretability results (global importance, bee swarm, and key variable plots). Section 4.6 concludes with the main takeaways and their implications for chapter 5.

4.1 Data exploration results (EDA)

Outcome distributions. The raw wellbeing score (mean of SATIS, WORTH, HAPPY) was right-skewed, with most respondents above 6, binarizing at 8 yielded near balanced classes (49% vs. 51%) (Figure 3.1, 3.2).

Redundancy pruning (correlation). Using Pearson correlations on the training set with a redundancy threshold of $|r| \geq 0.90$, we removed 147 of 470 candidate predictors (31.3%), retaining 323 (68.7%) for downstream selection. Pruning chiefly collapsed near-duplicates: tenure/furnishing indicators around `TIED` (e.g., `FURN`, `LLORD`, `TEN1`; $|r|$ up to 1.00); employment/ benefit-status families around `INECAC05` and `UCREDIT` (e.g., `ILODEFR`, `FTPTW`, `UNDEMP`, `JSATYP`; $|r| \approx 0.90$–0.98); hours-worked measures around `SUMHRS` (e.g., `TOTHRS`, `TTACHR`, `PAIDHRA`; $|r|$ up to 1.00); qualification/training flags around `CLAIMS14` (e.g., `QULNOW`, `ENROLL`, `APPR12`; $|r| \approx 0.99$–1.00); regional encodings around `NUTS163` (e.g., `ITL321`/`ITL221`/`GOR` codes; $|r| \approx 0.97$–1.00); illness day-of-week dummies around `ILLSUN` (e.g., `ILLMON`–`ILLFRI`;

$|r| \approx 1.00$); and paired health/limitation items (e.g., LIMITA with LIMITK = 1.00; DISEA with DISCURR20 = 0.98). Other singletons included REDACT with XDISDDA20 (0.94), LKSELA with LKTIMA/METHM (0.98), LIV12W with LIVTOG (1.00), SCHM12 with MF1664 (1.00), LOOKM111 with PREFHR (0.93), and QRTR with THIRQTR/THISQTR (1.00). The full pair-by-pair mapping is provided in Appendix C.

4.2 Feature selection results

NMI ranking (Fig 4.1). On the training data, ANXIOUS has NMI ≈ 0.060—about $3\times$ the next group ($\approx$0.018–0.019: REDACT, MARCHK, DISCURR20, LIMITK, LIMITA, MARDY6, DISEA, LIMACT). Absolute NMIs are modest (binary target, many discrete predictors), but the ordering is informative: Psychological distress leads, relationship status and activity limitation measures smaller, incremental signal.

Redundancy pruning (Fig 4.2). We removed pairs with Pearson $|r| \geq 0.90$h (training set), keeping the more interpretable/higher NMI representative. Key clusters include LIMITA-LIMITK, DISEA-DISCURR20, XDISDDA20-REDACT, tenure items around TIED (e.g., FURN/L-LORD/TEN1), and benefits/employment markers around UCREDIT. This stabilizes linear models, reduces variance in trees/boosting, and simplifies interpretation.

Top-K curve (Fig 4.3). A Random Forest probe shows rapid gains with the first few features, a small dip as noisier variables enter, then a broad plateau which is a classic diminishing returns. Balancing parsimony and accuracy, we fixed K = 7. In conclusion, the final 7 non-redandant features are **ANXIOUS, MARCHK, REDACT, MARDY6, LIMACT, LIMITK, DISCURR20**.

Reading the figures. In 4.1, labels on bars report exact NMI; in 4.2, "×" marks denote pruned pairs; in 4.3, the vertical axis is the mean of the three metrics (AUC, accuracy, F1), and the curve's early elbow motivates a compact K.

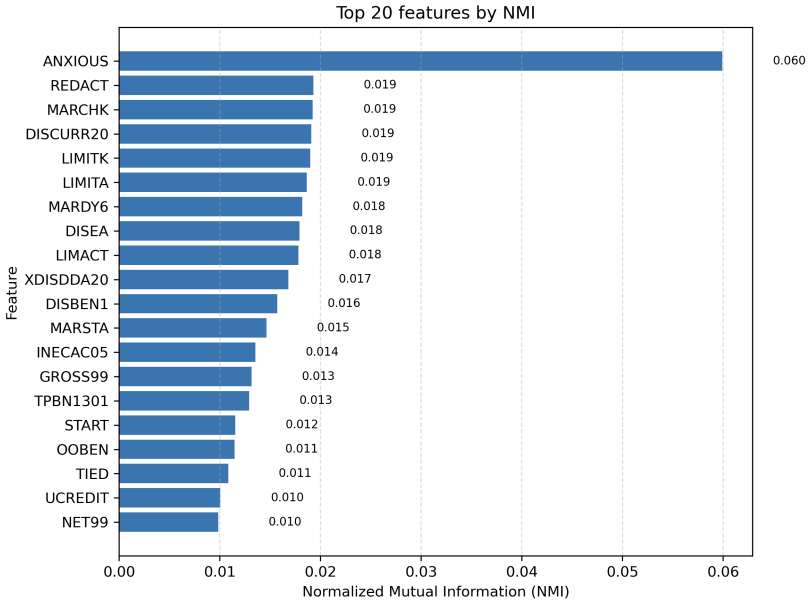

Figure 4.1: Top-20 NMI bar plot

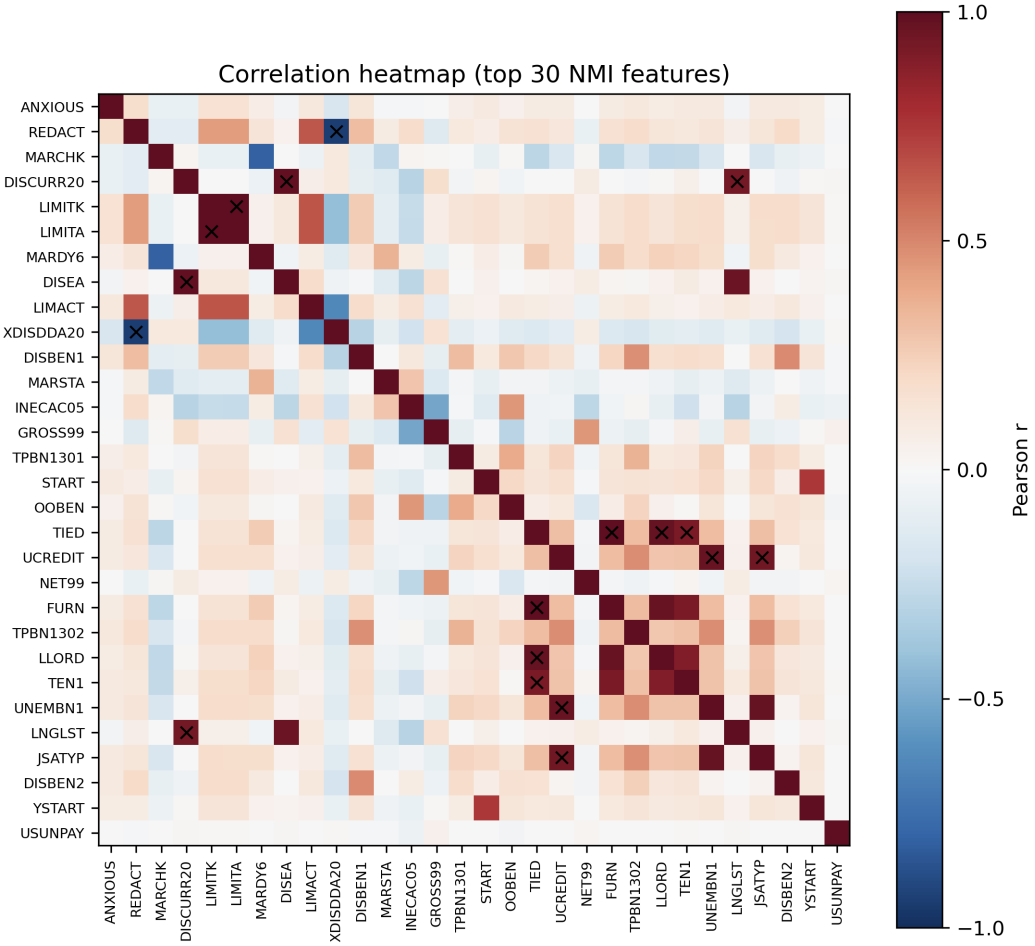

Figure 4.2: Correlation heat map with pruned pairs highlighted

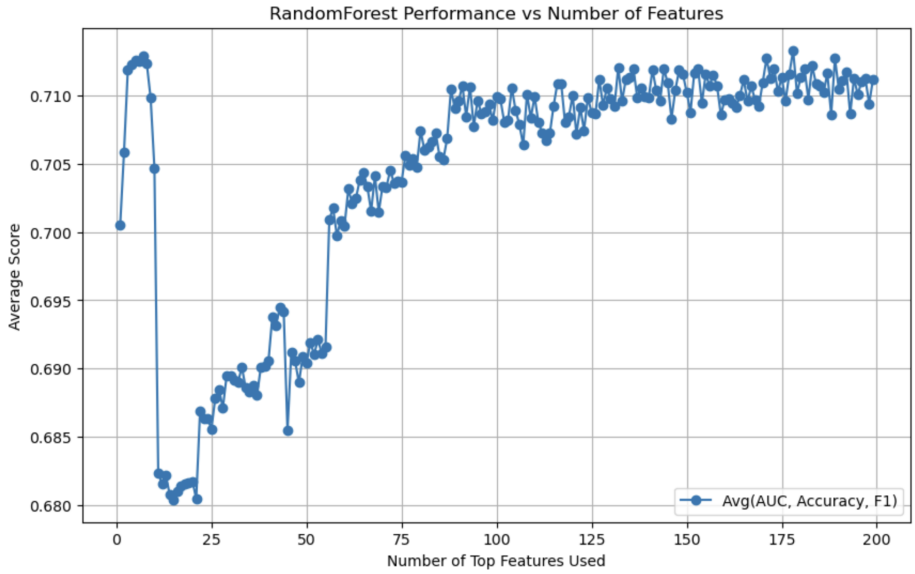

Figure 4.3: Performance vs. number of features (top-K curve)

4.3 Model Hyperparameter tuning

Consistent with 3.2.3, we tuned each model with GridSearchCV and stratified 3-fold CV on the training/validation data only. Below are the **best hyperparameter selected** by ROC–AUC (Accuracy and weighted F1 as checks). Each model was then **refit on train+validation** and evaluated once on the held out test set.

- **Logistic Regression:** C = 0.1, penalty = L2, solver = lbfgs
- **Ridge Classifier:** alpha = 0.1, solver = auto
- **Linear SVM:** C = 1.0, max_iter = 5000
- **KNN:** n_neighbors = 7, weights = uniform, p = 2
- **Random Forest:** n_estimators = 200, max_depth = 10, min_samples_split = 5, min_samples_leaf = 2
- **XGBoost:** n_estimators = 100, max_depth = 3, learning_rate = 0.1, subsample = 0.8, colsample_bytree = 1.0
- **MLP:** hidden_layer_sizes = (64, 32), activation = relu, learning_rate_init = 0.001, solver = adam, max_iter = 200

4.4 Model performance

The predictive framework developed in this study was designed to classify subjective wellbeing outcomes using a range of machine learning models. Model performance here refers to the ability of each algorithm to correctly distinguish between individuals reporting *Low-High* versus V*ery high* wellbeing, as measured by standard evaluation metrics (ROC-AUC, accuracy, F1-score).

The hyperparameter tuning stage demonstrated that ensemble tree based methods and neural networks provided the strongest predictive performance. Both XGBoost and Random Forest achieved the highest validation results, with ROC-AUC values of approximately 0.755, closely followed by the MLP Neural Network (ROC-AUC = 0.754). In contrast, linear models such as Ridge Classifier, Logistic Regression, and Linear SVM performed slightly lower (ROC-AUC $\approx$ 0.746), while KNN lagged behind with a ROC-AUC of 0.709. These results suggest that models capable of capturing non-linear feature interactions are better suited to the complexity of wellbeing prediction (Table 4.1).

On the independent test set, performance patterns were consistent with validation results. XGBoost achieved the strongest overall performance (Accuracy = 0.698, F1 = 0.698, ROC-AUC = 0.758). The MLP Neural Network performed almost identically (ROC-AUC = 0.757), while Random Forest also delivered strong predictive ability (ROC-AUC = 0.757). The linear models: Ridge Classifier, Logistic Regression, and Linear SVM maintained stable but lower results (ROC-AUC $\approx$ 0.747). Once again, KNN underperformed relative to other approaches (ROC-AUC = 0.712). These findings reinforce the conclusion that non-linear models consistently outperform linear baselines in predicting wellbeing outcomes (Table 4.2).

Table 4.1: Validation performance across models

Models	Accuracy	F1-score	ROC–AUC
XGBoost	**0.6939**	**0.6939**	**0.7554**
MLP Neural Network	0.6925	0.6926	0.7544
Random Forest	0.6939	0.6938	0.7548
Linear SVM	0.6885	0.6867	0.7466
Ridge Classifier	0.6885	0.6867	0.7465
Logistic Regression	0.6889	0.6873	0.7464
KNN	0.6660	0.6660	0.7091

Table 4.2: Final test set performance across models

Models	Accuracy	F1-score	ROC–AUC
XGBoost	**0.6978**	**0.6978**	**0.7577**
MLP Neural Network	0.6977	0.6977	0.7571
Random Forest	0.6978	0.6977	0.7569
Linear SVM	0.6894	0.6877	0.7470
Ridge Classifier	0.6894	0.6877	0.7470
Logistic Regression	0.6902	0.6888	0.7469
KNN	0.6708	0.6708	0.7125

Overall, the comparison demonstrates two key points: (1) non-linear models are more effective at capturing the multifactorial nature of wellbeing, and (2) performance differences, while modest, are consistent and robust across validation and test phases.

4.4.1 Confusion Matrix Analysis

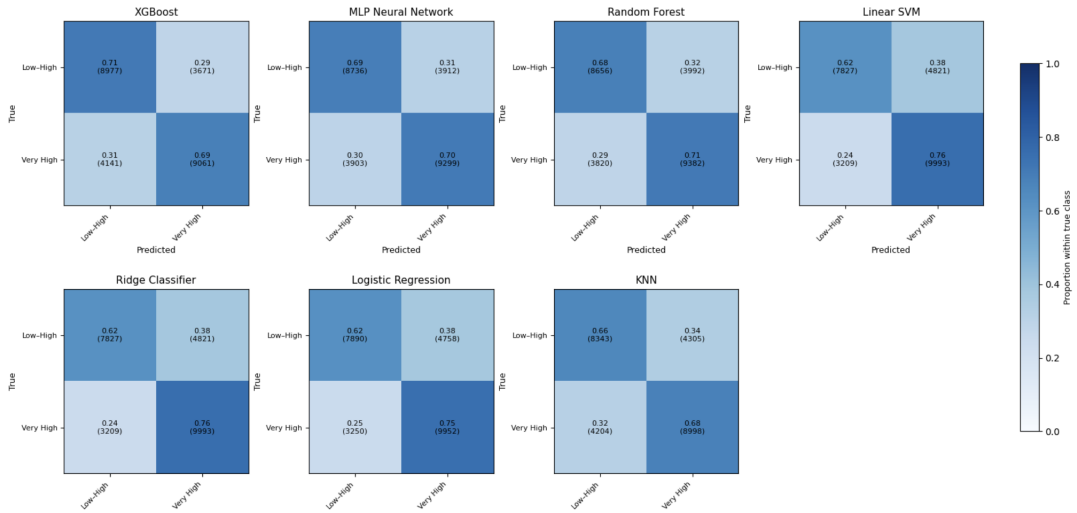

Figure 4.4: Multi-panel confusion matrices for all models (test set)

Across models, the non-linear methods perform best and most evenly. On the test set, XGBoost correctly classifies 8,977 Low–High and 9,061 Very High cases (off-diagonals 3,671/4,141), yielding row normalized diagonals of 0.71 vs 0.69 and ROC–AUC $\approx$ 0.758. MLP is similarly balanced with 8,736/9,299 correct (3,912/3,903 errors), diagonals 0.69 vs 0.70, AUC $\approx$ 0.757. Random Forest is close behind (8,656/9,382 correct; diagonals 0.68 vs 0.71; AUC $\approx$ 0.757). In contrast, the linear baselines show asymmetric recall: e.g., Logistic (7,890/9,952; 0.62 vs

0.75) and Linear SVM (7,827/9,993; 0.62 vs 0.76), producing lower AUCs ($\tilde{0}$.747). KNN is the weakest (8,343/8,998; 0.66 vs 0.68; AUC $\approx$ 0.713). Overall, XGBoost and MLP combine the highest discrimination with balanced diagonals (gaps $\approx$ 2 p.p.), whereas linear models exhibit a $\tilde{1}$4 p.p. gap favouring the *Very High* class, explaining their lower overall performance (Fig 4.4).

4.5 Model Interpretability

4.5.1 Global Feature Importance

To enhance interpretability, SHAP (SHAPley Additive Explanations) was applied to both XGBoost and the MLP Neural Network. For XGBoost, ANXIOUS emerged as the most influential feature (mean |SHAP| $\approx$ 0.80), far exceeding contributions from MARCHK (0.20), REDACT (0.17), MARDY6 (0.12), and physical limitation measures such as LIMACT and LIMITK (Fig 4.5).

The MLP Neural Network revealed a consistent pattern: ANXIOUS again dominated (mean |SHAP| $\approx$ 0.41), followed by REDACT, LIMACT, and LIMITK. The consistency across two distinct algorithmic families underscores the central role of psychological distress and physical health limitations in shaping wellbeing outcomes. The ordering shifts slightly between models, but the **pattern is stable.**

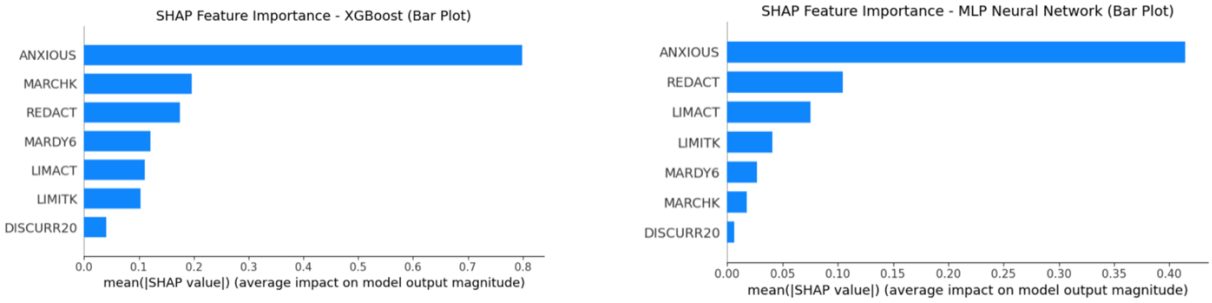

Figure 4.5: SHAP feature importance bar plots for XGBoost and MLP Neural Network

4.5.2 Local Explanations

The SHAP bee swarm plots (Fig 4.6) provide a finer grained perspective by showing how individual feature values influenced model predictions.

- **Higher anxiety scores** (ANXIOUS) were strongly associated with lower predicted wellbeing.
- **Fewer reported activity limitations** (LIMACT, LIMITK) corresponded to higher predicted wellbeing.

- **Employment related factors** (REDACT, MARCHK) further contributed, with positive employment states linked to improved wellbeing predictions.

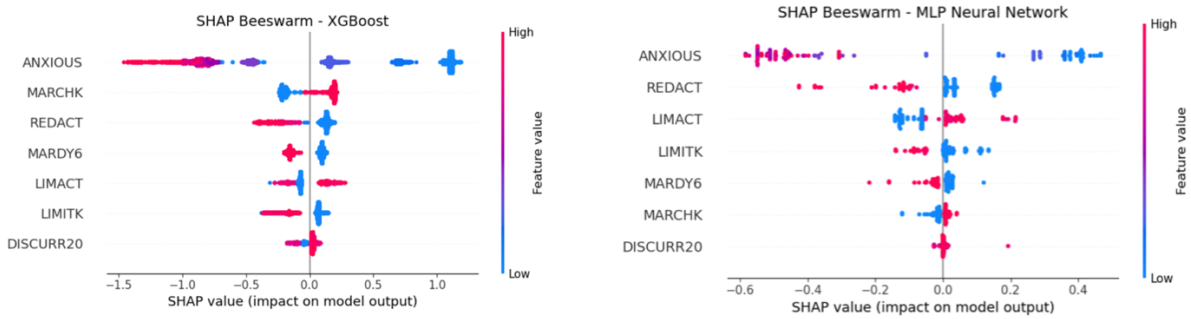

Figure 4.6: SHAP bee swarm plots for XGBoost and MLP Neural Network

These visualizations demonstrate that both psychological and physical health factors jointly SHAP wellbeing predictions. Importantly, they provide actionable insights by highlighting which variables most consistently shift model outputs.

4.5.3 SHAP Value Plots for Key Variables

To gain deeper insight into how categorical survey variables influenced predictions, SHAP value plots were generated for ANXIOUS, LIMACT, and MARDY6 (marital/partnership status) we can see Figure 4.7. **Anxiety (ANXIOUS):** Respondents who reported being "not at all anxious" substantially increased their likelihood of being classified as having "Very High" wellbeing. Conversely, high anxiety scores (8-10, "completely anxious") consistently shifted predictions toward the "Low-High" wellbeing category. **Activity limitation (LIMACT):** Individuals reporting no restrictions, or only minor limitations, were strongly associated with higher wellbeing predictions. Severe limitations exerted a significant negative effect. **Marital/partnership status (MARDY6):** Being married, cohabiting, or in a civil partnership increased predicted wellbeing, while respondents without a partner were more likely to be classified as lower wellbeing.

Collectively, these results reinforce established wellbeing research: psychological health, physical capacity, and social support are dominant determinants of subjective wellbeing. The models not only predicted outcomes accurately but also produced explanations that align with theoretical perspectives, enhancing their validity and practical utility.

4.6 Summary

This chapter shows that a compact specification can predict subjective wellbeing reliably. After correlation based redundancy pruning ($|r| \geq 0.90$), **147/470 (31.3%)** candidate predictors were removed, leaving **323** non-collinear variables. Using NMI ordering and a top-K probe, performance plateaued at $\mathbf{K = 7}$ final features: **ANXIOUS, MARCHK, REDACT, MARDY6, LIMACT, LIMITK, DISCURR20**.

Across models, non-linear methods dominated: **XGBoost**, a gradient boosted decision-tree

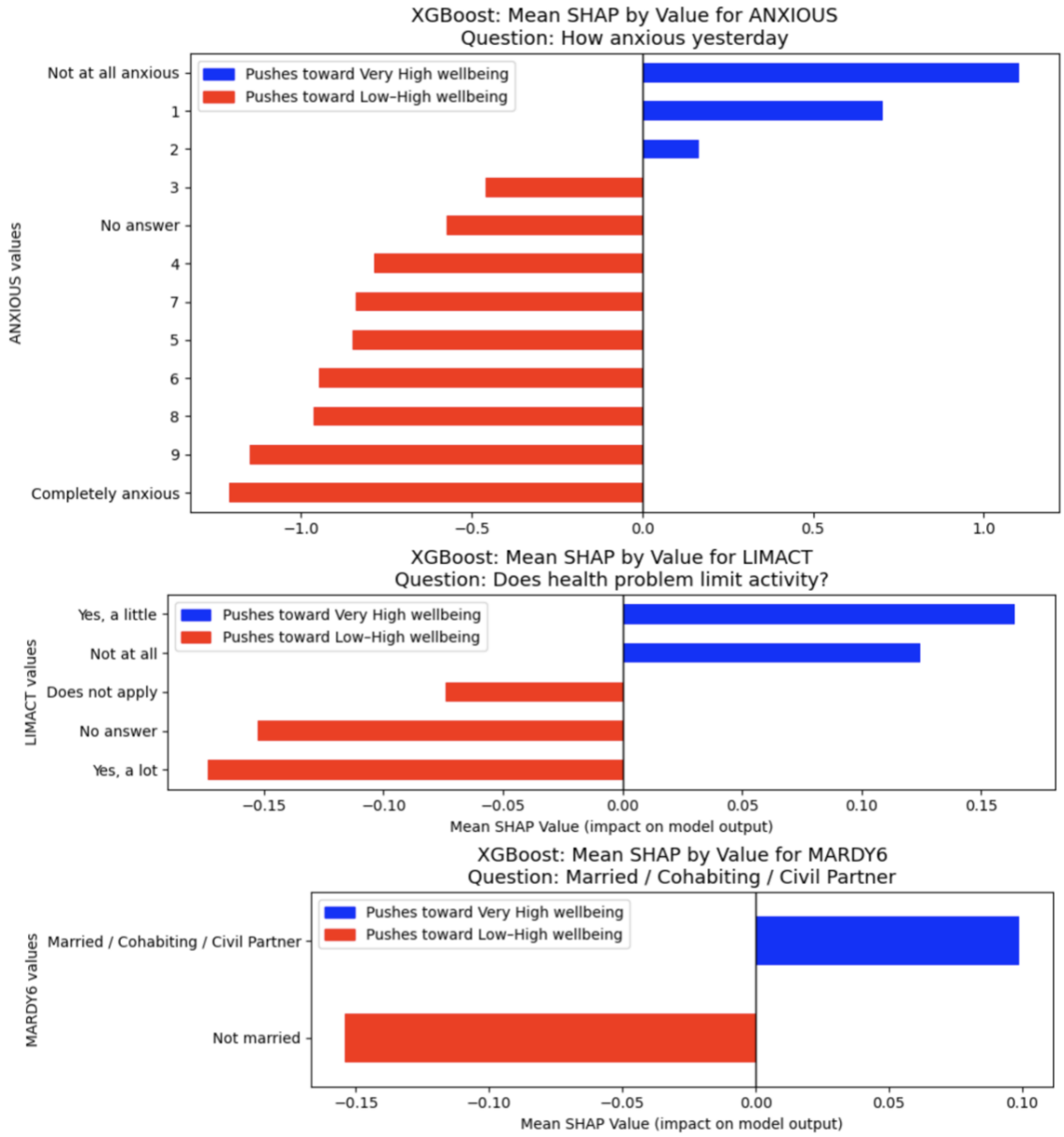

Figure 4.7: SHAP value plot for ANXIOUS, LIMACT and MARDY6 (XGBoost)

ensemble and **MLP**, a feed forward neural network delivered the best test discrimination (ROC–AUC = **0.758** and **0.757**, Accuracy/F1 ≈ **0.698**), with **Random Forest** close behind (AUC **0.757**). Linear baselines were slightly lower (AUC ≈**0.747**), and **KNN** underperformed (AUC **0.712**). Multi-panel confusion matrices show **balanced errors** across "Low-High" and "Very High," indicating no material class bias at the default cutoff. SHAP analyses are consistent across model families: **ANXIOUS** dominates, with functional limitations (**LIMACT, LIMITK**) and partnership/employment markers (**MARDY6, REDACT, MARCHK**) adding incremental signal. Overall, the seven feature model is **accurate, compact, and interpretable**, suitable for screening and targeted follow up, with thresholds adjustable to application needs.

Chapter 5

Discussion and Analysis

Motivation This chapter interprets the empirical findings and explains what they mean for wellbeing research and policy. It moves beyond reporting scores to ask why the best models (XGBoost and MLP) worked, which factors mattered most, and how a compact seven feature specification can deliver near state of the art accuracy while remaining interpretable. We test how stable the conclusions are under alternative setups (more features, different targets, stricter missing data handling), reflect on metric choices and model behaviour, and consider theoretical alignment with the wellbeing literature. Finally, we acknowledge limits and ethical guardrails so that any future use of these models remains fair, transparent, and proportionate.

Organization Section 5.1 interprets the main results and links SHAP patterns to the determinants of wellbeing. Section 5.2 assesses robustness (178 feature variant, alternative targets, missing code treatments) and shows the seven-feature model is the most efficient and stable. Section 5.3 discusses metric choices (AUC/F1 vs accuracy) and what they reveal about model behaviour. Section 5.4 situates the findings within theory and practice, highlighting where the models corroborate prior evidence and where they add nuance. Section 5.5 outlines limitations (performance ceiling, binary outcome, survey biases, SHAP approximations). Section 5.6 considers ethical, legal, and social implications, including fairness and GDPR aligned deployment. Section 5.7 closes with a concise summary to bridge to the conclusions in Chapter 6.

5.1 Interpretation of Results

The results demonstrate that the modelling framework successfully captured key determinants of wellbeing, while maintaining methodological rigour. Predictive performance across models was consistent, with non-linear models **XGBoost**, **Random Forest**, and the **MLP Neural Network** achieving the strongest results (Accuracy ≈ 0.70, AUC ≈ 0.75). These outcomes validate the modelling pipeline and align with prior wellbeing literature, which consistently identifies mental health (particularly anxiety levels), physical limitations, and social circumstances as the strongest correlates of self reported wellbeing.

This pattern is further illustrated in Figure 4.7, where SHAP categorical breakdowns show that lower anxiety, fewer reported limitations, and being in a marital or cohabiting partnership consistently shifted predictions toward "Very High" wellbeing. Importantly, the fact that predictive performance plateaued after the inclusion of approximately seven features demon-

strates that the most influential determinants of wellbeing can be distilled into a compact subset of variables without significant much loss in accuracy. This supports the robustness of the feature selection strategy, which combined Normalized Mutual Information, correlation pruning, and model based validation.

5.2 Robustness of Findings

To further test robustness, a re-estimated the whole model using the top 178 features instead of the top seven was evaluated. As shown in 5.1, the top performing models (XGBoost, Random Forest, and MLP) achieved almost identical outcomes across both setups, with a slight improvement in AUC ($\approx$ 0.76–0.77 vs. $\approx$ 0.75)

Table 5.1: Final test set performance across models (178 features).

Models	Accuracy	F1-score	ROC–AUC
XGBoost	**0.704487**	**0.704521**	**0.771705**
Random Forest	0.697563	0.697497	0.763653
MLP Neural Network	0.696789	0.696824	0.756552
Ridge Classifier	0.690909	0.689630	0.755026
Logistic Regression	0.556712	0.555817	0.588469
Linear SVM	0.533888	0.515289	0.550295
KNN	0.527621	0.527178	0.533040

This suggests that these algorithms are highly resilient to redundant predictors, supported by their internal mechanisms - regularization in XGBoost, feature bagging in Random Forest, and adaptive weight optimization in MLP

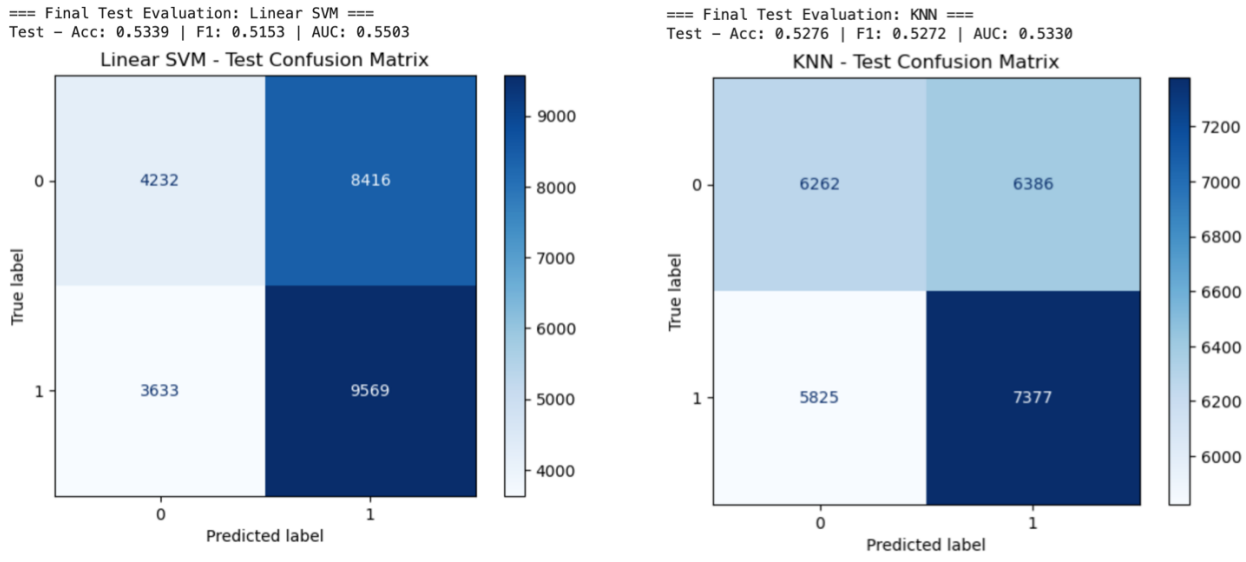

Figure 5.1: Confusion matrix for SVM and KNN (178 features)

By contrast, weaker models such as Logistic Regression, Linear SVM, and KNN performed considerably worse under the expanded feature set, with ROC-AUC scores dropping to near random levels ($\approx$0.53-0.59). Their collapse illustrates the curse of dimensionality and the absence of strong regularization or feature selection mechanisms. The error patterns are

illustrated in figure 5.1 ,which shows that high dimensional noise disproportionately increased misclassifications for these weaker algorithms.

Importantly, SHAP analyses (Fig 5.2) confirmed that the same core features ANXIOUS - dominant, LIMACT, and REDACT contributing: contribute predictions in both the 7 feature and 178 feature setups. This reinforces that the additional 171 attributes contributed little new signal.

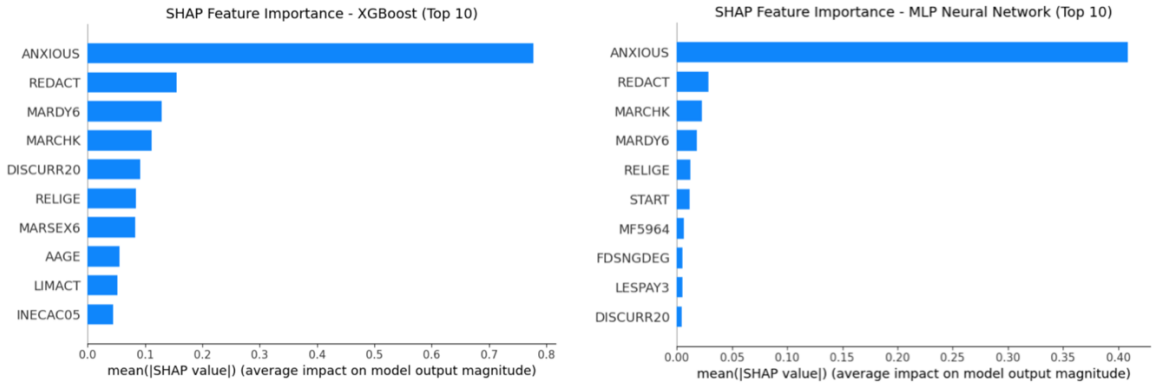

Figure 5.2: SHAP feature importance bar plots for XGBoost and MLP Neural Network (178 features)

Overall, while the 178-feature specification yields a small AUC gain for the strongest non-linear models, the seven-feature model is the more compact, efficient, and all-round choice: it retains nearly the same predictive strength, is easier to interpret, and avoids the instability seen in weaker algorithms as dimensionality rises.

Additional robustness trials validated this conclusion further. A binary split at 4 (Low wellbeing: <4, Medium-Very high: ≥ 4) produced superficially high accuracy (0.97) but was driven entirely by severe class imbalance, with models defaulting to the majority class. A 4 class categorization aligned with ONS thresholds (*Low, Medium, High and Very High*) achieved moderate accuracy ($\approx$0.59) and AUC ($\approx$0.76) but suffered from heavy confusion between adjacent categories, particularly between "Medium-High" and "Very High" wellbeing. In comparison, the binary split at 8 provided a balanced distribution ($\approx$49% vs. 51%), stable and consistent performance across algorithms, and the clearest trade off between predictive power and interpretability.

Beyond feature dimensionality and target categorization, data preprocessing itself was also stress tested. An additional robustness trial evaluated the effect of stricter handling of special missing codes. In this setup, all rows containing "-8" (No answer) were dropped, and columns with more than 20% "-9" (Does not apply) values were excluded, resulting in a dataset of 162,875 respondents and 93 attributes. Despite this substantial reduction, predictive performance remained highly consistent with the main pipeline (AUC $\approx$ 0.75 across top models). Results were slightly lower overall, suggesting that "-8" and "-9" codes may carry weak predictive signal, likely reflecting patterns in survey response behaviour or contextual applicability. This finding emphasizes a broader implication for survey based machine learning: aggressive exclusion of such codes does not improve model performance, and their retention may even provide marginal benefit. Nonetheless, the small effect sizes also indicate that missing code treatment is not a primary limitation, with performance more constrained

by the underlying complexity of wellbeing measurement.

5.3 Performance Metrics and Model Behaviour

In evaluating performance, **AUC** and **F1-score** were prioritized over raw accuracy. Accuracy can be misleading in cases of mild class imbalance, as it does not capture how well the model discriminates between "Low-High" and "Very High" wellbeing categories. AUC provides a threshold independent measure of separability, reflecting the model's ability to rank individuals correctly by wellbeing risk. Similarly, the weighted F1-score balances precision and recall, ensuring that both false positives and false negatives are penalized in a way that reflects real world relevance. These metrics therefore offered a more reliable assessment of predictive quality in this context than accuracy alone.

Among the tested algorithms, **XGBoost** and **MLP** consistently outperformed others. This can be attributed to their ability to capture complex, nonlinear relationships in the data. Wellbeing is influenced by subtle interactions between psychological, physical, and socio-economic variables, which linear models such as Logistic Regression or Ridge Classifier are unable to fully represent.

- **XGBoost** excels due to its ensemble learning framework, iterative boosting, and ability to handle feature interactions and non-linearities efficiently while controlling overfitting through regularization.
- **MLP Neural Network** leverages hidden layers and non-linear activation functions, enabling it to approximate complex mappings between predictors and wellbeing categories.
- By contrast, simpler models like **KNN** and **Linear SVM** were limited in their flexibility, leading to weaker performance.

Thus, the superior results of XGBoost and MLP demonstrate their suitability for modelling multifactorial constructs like wellbeing, where interactions between predictors are critical.

5.4 Theoretical and Practical Alignment

From a verification perspective, the project demonstrated best practices in data science. Data splitting into train, validation, and independent test sets ensured unbiased performance estimation and prevented data leakage. Hyperparameter tuning was performed systematically using stratified cross validation, and scaling was applied selectively to models requiring it (e.g., MLP).

Furthermore, model interpretability was addressed through SHAP analysis, which highlighted ANXIOUS, REDACT, and LIMACT as the most influential predictors. These findings resonate strongly with prior wellbeing research. Beyond feature ranking, the categorical SHAP plots helped bridge statistical modelling with sociological theory. They showed, for example, that psychological distress, physical capacity and marital status - long established in well-being research - manifested directly in model predictions, thus reinforcing both interpretability and theoretical alignment.

5.5 Limitations

At the same time, the study is not without limitations. The overall performance, while good, stabilized around an AUC of 0.75, suggesting that wellbeing is a complex construct with determinants beyond the scope of the dataset. The binary classification approach, while practical, may oversimplify the nuanced spectrum of wellbeing. Moreover, reliance on self reported survey data introduces potential biases, and the Kernel SHAP approach for MLP required subsampling due to computational constraints, which may limit the precision of interpretability outputs. Finally, subgroup fairness was not assessed: we did not audit performance or calibration across key demographics (e.g., age, sex, region), so potential disparities remain unquantified. Despite these challenges, the project contributes meaningfully by balancing predictive accuracy, methodological soundness, and interpretability. It provides a framework that can be extended to more granular wellbeing categories, additional predictors, or alternative modelling techniques in future research.

5.6 Ethical, Legal, and Social Dicussion

While this study was conducted using anonymized secondary data in full compliance with UK Data Service licensing conditions and the principles of the General Data Protection Regulation (GDPR), several broader ethical and social issues warrant consideration. The use of survey responses for predictive modelling raises questions of consent, as participants were not explicitly informed that their data might be employed for wellbeing prediction. Furthermore, even accurate models could be misapplied in sensitive contexts such as insurance assessment or targeted interventions, where individuals with lower predicted wellbeing risk being stigmatized.

Concerns about bias and fairness also remain. Although the models achieved balanced performance across the two wellbeing categories, the study did not examine subgroup level disparities (for example, by age, gender, or ethnicity). Future research should therefore include fairness assessments to ensure that predictive approaches do not inadvertently reinforce existing social inequalities. From a legal standpoint, any use of such models outside academia would require a lawful basis for data processing, the completion of Data Protection Impact Assessments (DPIAs), and strict adherence to GDPR requirements concerning transparency, proportionality, and data minimization. These measures were beyond the scope of the present project but are essential for any real world application.

Overall, while the study makes a methodological and substantive contribution to wellbeing research, its findings should be interpreted with caution. Any applied use of predictive modelling in this area must be carefully governed to protect privacy, ensure fairness, and maintain public trust.

5.7 Summary

This chapter has discussed the results of wellbeing prediction models, emphasizing both their predictive performance and interpretability. Robustness analyses confirmed that a binary split at 8 and a compact seven feature framework provided the most efficient and stable solution, while SHAP analyses consistently highlighted anxiety as dominant predictors.

The significance of these findings lies in their ability to replicate known associations from wellbeing research while quantifying their predictive power using modern machine learning. Although limitations remain, the project demonstrates that interpretable, data driven models can provide actionable insights for both academic research and policy making. The next chapter concludes the study by summarizing its contributions and outlining avenues for future work.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This project investigated whether machine learning could provide robust and interpretable models of subjective wellbeing in the UK, using nationally representative survey data. The aim was to predict wellbeing outcomes with high accuracy while also identifying the minimal set of predictors required for strong performance. To achieve this, the study set five objectives: **(1)** prepare and preprocess the Annual Population Survey dataset, **(2)** compare multiple algorithms spanning linear, ensemble, and neural approaches, **(3)** apply systematic hyperparameter tuning, **(4)** evaluate performance using robust metrics, and **(5)** interpret model behaviour with explainable AI techniques.

All objectives were successfully met. The modelling pipeline demonstrated that wellbeing prediction is feasible with good accuracy: the strongest models, **XGBoost**, **Random Forest**, and the **MLP Neural Network**, consistently achieved ROC-AUC values around 0.75 and accuracies near 70%. These results confirm that non-linear models capture complex relationships between wellbeing and its determinants more effectively than linear baselines such as Logistic Regression or SVM.

A key achievement was the discovery that predictive performance plateaued after the inclusion of only **seven non-redundant features**. This finding, derived through Normalized Mutual Information, correlation pruning, and validation with Random Forest, shows that compact and efficient models are possible without sacrificing performance. **SHAP analysis** further revealed that psychological distress (ANXIOUS), physical limitations (LIMACT, LIMITK), and relationship status (MARDY6) were consistently the strongest predictors of wellbeing. These results are highly consistent with existing wellbeing literature, which identifies mental health, physical health, and social ties as core drivers of life satisfaction and happiness.

Robustness checks confirmed the stability of these findings. Expanding to 178 features did not meaningfully improve performance, while alternative target splits (binary at 4, or 4 class categorization) reduced interpretability and stability. Stricter handling of missing codes slightly lowered results but reinforced that the modelling pipeline was not overly sensitive to data cleaning decisions. Collectively, these trials confirm that the **binary split at 8 with seven features** offers the most reliable, interpretable, and efficient framework for wellbeing prediction.

Overall, the project contributes to the field by demonstrating that advanced machine learning can be applied responsibly to wellbeing research, balancing predictive strength with interpretability. The results have practical implications for **survey design**, by suggesting that shorter and more focused wellbeing assessments may still yield robust predictive insights, and for **policy analysis**, by highlighting anxiety, health, and relational factors as the strongest levers for intervention.

6.2 Originality of Contribution

This project makes several original contributions to wellbeing research:

1. **Feature efficient modelling**: Unlike most large scale wellbeing studies that retain hundreds of predictors, this study systematically reduced the predictor set from 471 to just seven variables without losing predictive power (AUC $\approx$ 0.75). This offers a novel pathway toward leaner, more cost effective surveys that maintain accuracy while reducing respondent burden.
2. **Integration of predictive strength with interpretability**: By applying advanced models (XGBoost and MLP neural networks) alongside SHAP explainability tools, the study bridged the gap between high performance and transparency. It demonstrated that complex "black-box" models can be rendered interpretable in both global and individual terms, addressing a persistent limitation in wellbeing research.
3. **Systematic robustness testing**: Through comparative trials with expanded feature sets, alternative classification schemes, and stricter data cleaning, the project explicitly tested the stability of its framework. This systematic validation is rarely undertaken in wellbeing analytics and strengthens the credibility of the conclusions.

Taken together, these contributions advance the methodological toolkit of wellbeing science, showing that compact, interpretable, and robust machine learning models can complement and in some cases outperform traditional regression approaches.

6.3 Future Work

While the project achieved its objectives, several opportunities remain for future research.

Data and scope. The APS provides rich survey data but is cross sectional and self reported. Future work should explore **prospective data datasets** to capture wellbeing trajectories over time, improving causal inference. Integration of **objective data sources** such as electronic health records, physical activity trackers, or social media behaviour could further enrich predictive capacity and validate self-reported measures.

Modelling approaches. Although ensemble methods and MLP proved effective, further experimentation with **deep learning architectures** such as transformers or graph neural networks may capture latent structures in wellbeing data more effectively. At the same time, causal modelling techniques could move beyond prediction to explore directional influences of

wellbeing determinants. Ensuring interpretability remains central, future research should also test emerging explainability frameworks alongside SHAP.

Outcome modelling. This study focused on a binary classification split at 8, which balanced interpretability and stability. Future research could revisit **multi class categorizations** aligned with ONS thresholds or treat wellbeing as a **continuous regression target**, providing more powerful insights into the spectrum of life satisfaction. Hybrid frameworks combining classification and regression may also prove valuable.

Fairness and generalization. Future work should assess whether predictive performance is consistent across demographic subgroups for example age, gender, ethnicity. Detecting and correcting biases is essential for equitable applications. Testing the framework in **international contexts** would also help determine whether predictors of wellbeing generalize beyond the UK.

Drift monitoring and model lifecycle. For applied use, the model should be monitored over time for **covariate drift** (inputs shifting), **label/prior drift** (class balance changing), and **concept drift** (changing relationships between features and wellbeing). Practical steps include tracking distributional shifts (e.g., PSI/KS tests), periodic performance and calibration checks (AUC, F1, Brier/ECE), SHAP-rank stability, and defining retraining/recalibration triggers and governance (versioning, model cards, audit logs).

Practical applications. For the models to be useful in real world decision making, they must be validated in applied settings. Future work could involve developing **policy facing dashboards** or decision support systems that highlight wellbeing risk factors in a transparent manner. Strong governance frameworks, ethical oversight, and compliance with regulations such as GDPR will be essential to prevent misuse and to build public trust.

In summary, future work should extend this study by incorporating higher dimension and more complex data, exploring advanced algorithms, refining outcome modelling, addressing fairness and generality, drift monitoring and testing applied implementations. These directions would not only deepen scientific understanding of wellbeing but also support its translation into evidence based policy and public health practice.

Chapter 7

Reflection

Completing this project has been an intensive learning experience that combined both academic understanding and technical skill development. On the academic side, I came to appreciate the complexity of wellbeing as a construct. It quickly became clear that wellbeing is not shaped by any single determinant but by the interplay of psychological, physical, and social factors. Translating this complexity into a predictive framework required careful methodological choices. I learned the importance of handling missing data thoughtfully, weighing the trade offs between dropping variables and applying imputation strategies such as mean replacement. Similarly, I developed a deeper understanding of feature selection techniques, particularly the use of Normalized Mutual Information (NMI) and correlation penalties to avoid redundancy. Exploring the strengths and weaknesses of different models was also invaluable: while tree based methods and neural networks captured complex relationships effectively, simpler linear models provided transparency but struggled with non-linear interactions. These experiences reinforced that methodological rigour requires balancing predictive strength with interpretability and practical considerations.

On the technical side, the project provided significant exposure to working with large scale survey data. Managing a dataset of over 340,000 respondents and hundreds of attributes required not only computational efficiency but also an appreciation of how different variable types demand different preprocessing. For instance, binary indicators, categorical codes, and continuous variables each required distinct treatment to preserve their meaning. I also became more confident in model development pipelines, from splitting into train/validation/test sets, to retraining tuned models on combined train + validation data, and finally running them on the test set. The use of class balancing strategies and SHAP explainability were both new techniques to me, and mastering them was crucial for handling imbalanced data and validating "black-box" models.

The project also presented several challenges. Initially, my concept and direction were heavily influenced by the ONS framework, which I later realized did not work well for my dataset. This forced me to restart from the ground up, rebuilding my approach around feature selection rather than relying on predefined attribute sets. Another major challenge was class imbalance, even after applying balancing strategies, the minority classes in alternative formulations remained difficult to predict. I also encountered the inherent limitations of self reported survey data, which introduces biases that cannot be fully corrected. Finally, applying Kernel SHAP to neural networks created computational constraints, requiring subsampling that reduced preci-

sion. These challenges taught me the importance of flexibility, critical thinking, and resource management in data science projects.

Decision making was another key part of the learning experience. I chose a binary split at 8 because it removed class imbalance while preserving interpretability, making the models more stable. For interpretability, I selected SHAP because it is both widely used and relatively simple to apply, while recognizing that more complex XAI approaches could be explored in future research. These choices highlighted how methodological decisions must balance theoretical justification, computational feasibility, and transparency.

From a personal perspective, the project fostered substantial growth. Being able to design and execute a complete pipeline has made me more confident in my ability to handle large and complex datasets in future work, and it motivates me to explore different modelling approaches. If I were to repeat the project, I would proceed more step by step, ensuring each stage of the pipeline was robust before moving forward, while still remaining open to new possibilities and ready to rebuild when needed. This mindset — balancing thorough planning with adaptability is one of the most valuable lessons I take away.

Overall, the project not only strengthened my technical and academic skillset but also enhanced my confidence in applying data science to real world social and health challenges. It has shown me that rigorous methodology, transparent interpretation, and resilience in the face of setbacks are just as important as technical skill, and these lessons will guide both my future research and professional practice.

References

Abdul Rahman, H., Kwicklis, M., Ottom, M., Amornsriwatanakul, A., Abdul-Mumin, K. H., Rosenberg, M. and Dinov, I. D. (2023), 'Machine learning-based prediction of mental well-being using health behavior data from university students', *Bioengineering* **10**(5), 575.
URL: *https://doi.org/10.3390/bioengineering10050575*

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**, 5–32.
URL: *https://link.springer.com/article/10.1023/A:1010933404324*

Centre for Data Ethics and Innovation (2020), Review into bias in algorithmic decision-making, Technical report, CDEI.
URL: *https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making*

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002), 'Smote: Synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research* **16**, 321–357.
URL: *https://www.jair.org/index.php/jair/article/view/10302*

Chen, T. and Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)', ACM, pp. 785–794.
URL: *https://dl.acm.org/doi/10.1145/2939672.2939785*

Clark, A. E., Flèche, S., Layard, R., Powdthavee, N. and Ward, G. (2018), *The Origins of Happiness: The Science of Well-Being over the Life Course*, Princeton University Press, Princeton, NJ. Sample chapter: https://assets.press.princeton.edu/chapters/i11179.pdf.
URL: *https://books.google.com/books?id=mHaODwAAQBAJ*

Collins, G. S., Reitsma, J. B., Altman, D. G. and Moons, K. G. M. (2015), 'Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement', *BMJ* **350**, g7594.
URL: *https://pubmed.ncbi.nlm.nih.gov/25569120/*

Cortes, C. and Vapnik, V. (1995), 'Support-vector networks', *Machine Learning* **20**(3), 273–297.
URL: *https://link.springer.com/article/10.1007/BF00994018*

Cover, T. M. and Hart, P. E. (1967), 'Nearest neighbor pattern classification', *IEEE Transactions on Information Theory* **13**(1), 21–27.
URL: *https://ieeexplore.ieee.org/document/1053964*

Diener, E., Lucas, R. E. and Oishi, S. (2018), 'Advances and open questions in the science of subjective well-being', *Collabra: Psychology* **4**(1), 15.
URL: *https://pmc.ncbi.nlm.nih.gov/articles/PMC6329388/*

Fawcett, T. (2006), 'An introduction to roc analysis', *Pattern Recognition Letters* **27**(8), 861–874.
URL: *https://www.sciencedirect.com/science/article/pii/S016786550500303X*

Ferrer-i Carbonell, A. and Frijters, P. (2004), 'How important is methodology for the estimates of the determinants of happiness?', *The Economic Journal* **114**(497), 641–659.
URL: *https://doi.org/10.1111/j.1468-0297.2004.00235.x*

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. and Bouchachia, A. (2014), 'A survey on concept drift adaptation', *ACM Computing Surveys* **46**(4), 44:1–44:37.
URL: *https://dl.acm.org/doi/10.1145/2523813*

Goodfellow, I., Bengio, Y. and Courville, A. (2016), *Deep Learning*, MIT Press, Cambridge, MA.
URL: *https://www.deeplearningbook.org/*

Helliwell, J. F. and Putnam, R. D. (2004), 'The social context of well-being', *Philosophical Transactions of the Royal Society B: Biological Sciences* **359**(1449), 1435–1446. Open access: https://pmc.ncbi.nlm.nih.gov/articles/PMC1693420/.
URL: *https://royalsocietypublishing.org/doi/10.1098/rstb.2004.1522*

HM Treasury (2021), Green book supplementary guidance: wellbeing, Technical report, HM Treasury.
URL: *https://www.gov.uk/government/publications/green-book-supplementary-guidance-wellbeing*

Information Commissioner's Office (2023), Data protection impact assessments (dpias): guide to accountability and governance, Technical report.
URL: *https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/guide-to-accountability-and-governance/data-protection-impact-assessments/*

Islam, M. R., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H. and Ulhaq, A. (2018), 'Depression detection from social network data using machine learning techniques', *Health Information Science and Systems* **6**(1), 8.
URL: *https://doi.org/10.1007/s13755-018-0046-0*

ISPOR (2025), 'Value in Health Journal Spotlights "Whole Health" in Special Series'. ISPOR News; press release.
URL: *https://www.ispor.org/heor-resources/news-top/news/2025/05/19/value-in-health-journal-spotlights--whole-health--in-special-series*

Kilfoyle, M. (2023), 'How is the cost of living crisis affecting public health?', *Economics Observatory* .
URL: *https://www.economicsobservatory.com/how-is-the-cost-of-living-crisis-affecting-*

King's College London (2023), 'Cost-of-living crisis is worsening the mental health of most vulnerable'.
URL: *https://www.kcl.ac.uk/news/cost-of-living-crisis-is-worsening-the-mental-health*

Kohavi, R. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, *in* 'Proceedings of IJCAI-95'.
URL: *https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf*

Kuhn, M. (2008), 'Building predictive models in r using the caret package', *Journal of Statistical Software* **28**(5), 1–26.
URL: *https://www.jstatsoft.org/v28/i05/*

Liu, Y. (2024), 'Multilayer perceptron-based literature reading preferences predict anxiety and depression in university students', *Frontiers in Psychology* **15**, 1425471.
URL: *https://doi.org/10.3389/fpsyg.2024.1425471*

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.-I. (2020), 'From local explanations to global understanding with explainable AI for trees', *Nature Machine Intelligence* **2**, 56–67.
URL: *https://doi.org/10.1038/s42256-019-0138-9*

Lundberg, S. M. and Lee, S.-I. (2017*a*), A unified approach to interpreting model predictions, *in* 'Advances in Neural Information Processing Systems 30 (NeurIPS 2017)', pp. 4765–4774.
URL: *https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf*

Lundberg, S. M. and Lee, S.-I. (2017*b*), A unified approach to interpreting model predictions, *in* 'Advances in Neural Information Processing Systems 30 (NeurIPS 2017)', Curran Associates, Inc., pp. 4765–4774.
URL: *https://doi.org/10.48550/arXiv.1705.07874*

Mind (2025), 'Cost of living crisis and mental health'.

Office for National Statistics (2022), Personal well-being in the uk: April 2021 to march 2022, Technical report, Office for National Statistics, Newport, UK. Statistical bulletin.
URL: *https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/measuringnationalwellbeing/april2021tomarch2022*

Office for National Statistics (2023), Personal well-being in the uk: April 2022 to march 2023, Technical report, Office for National Statistics, Newport, UK. Statistical bulletin.
URL: *https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/measuringnationalwellbeing/april2022tomarch2023*

Office for National Statistics (2024), 'Annual population survey three-year pooled dataset, january 2021–december 2023'. Study number: 9291. Data collection distributed July 2024.
URL: *https://doc.ukdataservice.ac.uk/doc/9291/mrdoc/UKDA/UKDA_Study_9291_Information.htm*

Office for National Statistics (2025), Public opinions and social trends, great britain: April 2025, Statistical bulletin, Office for National Statistics. Release date: 16 May 2025.
URL: *https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/publicopinionsandsocialtrendsgreatbritain/april2025*

Oparina, E., Kaiser, C., Gentile, N., Tkatchenko, A., Clark, A. E., De Neve, J.-E. and D'Ambrosio, C. (2022), 'Human wellbeing and machine learning'.
URL: *https://arxiv.org/abs/2206.00574*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É. (2011), 'Scikit-learn: Machine learning in python', *Journal of Machine Learning Research* **12**, 2825–2830.
URL: *https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf*

Powdthavee, N. (2008), 'Putting a price tag on friends, relatives, and neighbours: Using surveys of life satisfaction to value social relationships', *The Journal of Socio-Economics* **37**(4), 1459–1480. Publisher page: https://www.sciencedirect.com/science/article/pii/S1053535707001205.
URL: *https://ideas.repec.org/a/eee/soceco/v37y2008i4p1459-1480.html*

Sokolova, M. and Lapalme, G. (2009), 'A systematic analysis of performance measures for classification tasks', *Information Processing & Management* **45**(4), 427–437.
URL: *https://www.sciencedirect.com/science/article/pii/S0306457309000259*

Steptoe, A., Deaton, A. and Stone, A. A. (2015), 'Subjective wellbeing, health, and ageing', *The Lancet* **385**(9968), 640–648.
URL: *https://pubmed.ncbi.nlm.nih.gov/25468152/*

Supranowicz, P. and Małgorzata, P. (2014), 'Holistic measurement of well-being: psychometric properties of the physical, mental and social well-being scale (pmsw-21) for adults', *Rocz Panstw Zakl Hig* **65**(3), 251–258.
URL: *https://pubmed.ncbi.nlm.nih.gov/25247806/*

Tapper, J., Fazackerley, A. and Thorpe, V. (2025), "'the pandemic reinforced existing inequalities – it was a magnifying glass": how Covid changed britain', *The Guardian* .
URL: *https://www.theguardian.com/world/2025/mar/09/the-pandemic-reinforced-existing-inequalities-it-was-a-magnifying-glass-how-covid-chan*

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., Parkinson, J., Secker, J. and Stewart-Brown, S. (2007), 'The warwick-edinburgh mental well-being scale (wemwbs): development and uk validation', *Health and Quality of Life Outcomes* **5**, 63.
URL: *https://hqlo.biomedcentral.com/articles/10.1186/1477-7525-5-63*

Topp, C. W., Østergaard, S. D., Søndergaard, S. and Bech, P. (2015), 'The who-5 well-being index: a systematic review of the literature', *Psychotherapy and Psychosomatics* **84**(3), 167–176.
URL: *https://pubmed.ncbi.nlm.nih.gov/25831962/*

UK Data Service (2024), 'End user licence agreement'.
URL: *https://ukdataservice.ac.uk/app/uploads/cd137-enduserlicence.pdf*

Varoquaux, G. (2018), 'Cross-validation failure: Small sample sizes lead to large error bars', *NeuroImage* **180**, 68–77.
URL: *https://pubmed.ncbi.nlm.nih.gov/28655633/*

Vassilev, G., Manclossi, S., Pyle, E., Parmar, M., Yull, J., Sidhu, S., Payne, C. and Tabor, D. (2019), Personal and economic well-being: what matters most to our life satisfaction?, Article, Office for National Statistics. Release date: 15 May 2019.
URL: *https://www.ons.gov.uk/peoplepopulationandcommunity/*

wellbeing/articles/personalandeconomicwellbeingintheuk/whatmattersmosttoourlifesatisfaction

Vinh, N. X., Epps, J. and Bailey, J. (2010), 'Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance', *Journal of Machine Learning Research* **11**, 2837–2854.
URL: *https://jmlr.csail.mit.edu/papers/volume11/vinh10a/vinh10a.pdf*

Ware, J. E. J. and Sherbourne, C. D. (1992), 'The mos 36-item short-form health survey (sf-36). i. conceptual framework and item selection', *Medical Care* **30**(6), 473–483.
URL: *https://pubmed.ncbi.nlm.nih.gov/1593914/*

World Health Organization (1948), 'Preamble to the constitution of the world health organization', Official Records of the World Health Organization, no. 2, p. 100.

Appendix A

Project Specification Form

MSc Project Specification Form

Version 03

Please complete this form in printed or typed text (Times New Roman size 11). A copy of the approved document will have to be included as appendix in the actual dissertation to help establish to which extend the project has been successful.

Section A

1. Project identification

Proposed dissertation title (maximum 15 words)	
Prediction for personal wellness	

2. Student details

Name (*in full*)
Yiu Ting Wong
e-mail address vg834361@student.reading.ac.uk

Term time address
St Georges Hall, Upper Redlands Road, Reading, UK RG1 5HZ
Contact telephone number

3. Supervisor
Name and contact details of the staff member

Name Ferran Espuny-Pujol
e-mail address f.espuny-pujol@reading.ac.uk

4. The supervisor *The person identified in Section A4 hereby approves the dissertation specification*

Name

Signature

Date

Fill in section A5 and A6, if the project has industrial input.

5. Company Partner
Name of organisation involved in the project

6. *Details of contact person in the organisation involved in the project*

Title *(e.g. Mr/Mrs/Dr)*	Name	
Address		
Tel. No.	Fax No.	e-mail address

MSc Project Specification Form

Section B – Overall Programme

1. Background and Literature review

Please describe in the space provided the background to the project and write a short literature review highlighting relevant developments on the topic of the dissertation

Personal well-being has become a crucial focus for public policy, especially in the wake of the COVID-19 pandemic. According to the UK Office for National Statistics (ONS), average ratings for personal well-being measures including life satisfaction, feelings of worthwhileness, happiness, and anxiety declined significantly in the years following the pandemic. An ONS analysis using the UK Annual Population Survey (APS) showed that self-reported health and marital status were key contributors to life satisfaction between May 2022 and May 2023.

Existing research typically applies linear regression methods to model well-being outcomes based on demographic and health-related variables. While useful, these models may oversimplify complex interrelationships between predictors. This project builds on ONS's analysis by incorporating additional risk factors and applying machine learning techniques (e.g., regularised regression, decision trees) to improve predictive accuracy. This work is made possible by accessing the APS dataset through the UK Data Service under the End User Licence.

Background on Well-being Determinants

A sizable amount of research has investigated the social, demographic, and economic factors that influence personal well-being. The Office for National Statistics (ONS) has consistently identified self-reported health, marital status, economic activity, and age as the strongest predictors of life satisfaction in the UK (ONS, 2023). Additional support from the ONS (2019) suggests that how individuals spend their income, particularly on experiences rather than necessities can significantly influence satisfaction levels. Academic studies further reinforce these findings for example, Powdthavee (2008) demonstrated that social relationships and household circumstances can be valued in monetary terms due to their large effects on life satisfaction. Similarly, Clark and Oswald (2002) found that factors like unemployment, marital breakdown, and poor health exert significant downward pressure on subjective well-being. To determinant of well being is different for everyone and one goal of this research is to find out what factors affect the wellbeing level of UK citizens the most.

Predictive Modelling in Well-being Research

While traditional regression models are widely used in well-being analysis, recent research has explored machine learning as an alternative. Nguyen and Li (2021) compared linear regression to decision trees and ensemble methods in predicting life satisfaction using socio-demographic data, finding that tree-based methods often yield better performance. Fernandes et al. (2020) applied similar techniques in health-related quality of life prediction, demonstrating that models such as random forests and gradient boosting not only improve accuracy but also offer practical value in identifying at-risk individuals. To enhance transparency in these models, SHAP (SHapley Additive exPlanations) has emerged as a popular tool for interpreting feature importance and individual predictions, especially in complex, non-linear models (Lundberg and Lee, 2017). By quantifying each variable's contribution to the output, SHAP makes it possible to uncover meaningful relationships that might remain hidden in traditional linear approaches. These developments support a shift towards more accurate and interpretable predictive analytics in public health and well-being research, particularly when working with rich datasets like the UK Annual Population Survey.

Model Interpretability and Policy Relevance

While machine learning methods offer improved predictive power, their practical value depends heavily on their interpretability — especially when informing public policy. In well-being research, it is important not only to make accurate predictions but also to understand which variables are driving those outcomes. Tools like SHAP (SHapley Additive exPlanations) are increasingly used to interpret complex models by attributing a clear, additive contribution of each input feature to a given prediction. This level of transparency is critical in models applied to social data, as it allows researchers and policymakers to validate findings, detect potential biases, and communicate results in an actionable way. By combining predictive accuracy with interpretability, the project aims to bridge the gap between advanced analytics and the real-world policy applications of well-being modelling.

MSc Project Specification Form

Section B – Overall Programme (continued)

2. Research question, justification and objectives

Please describe in the space provided the research question to be answered, justify why the topic is important at the present time, and describe the specific objectives of research against which your achievements will be measured.

Research Question: To what extent can predictive models perform better than the baseline linear regression approach used by the Office for National Statistics (ONS) and through transparent models what are the factors that contribute to the wellbeing of UK citizens the most.

Well-being is becoming a key area for government and policy decisions, especially since levels of life satisfaction have dropped in the UK during and after the pandemic. The ONS has used simple statistical models to understand what are the factor that affects people's happiness and satisfaction, mainly pointing to health and relationships. But these models can miss more complex patterns in the data. This research will use more advanced machine learning methods that discovers deeper insights, while still being explainable. The goal is not just to build more accurate models, but also to understand clearly what drives wellbeing so that these insights can be applied in real world decisions.

Objectives

1. Rebuild the ONS model with the same dataset
2. Select more useful and modern features to predict wellbeing
3. Predict result with different models like RF, gradient boost
4. Explain which variables are important with tools like SHAP
5. Compare performance of different models

MSc Project Specification Form

Section B – Overall Programme (continued)

3. Methodology

Please describe the methodology that you will use to achieve the objectives stated in Section B2.

You should also identify the data availability and quality, technical challenges, resource constraints, any other risks associate with the project and your plan to manage them.

This project will use data from the UK Annual Population Survey, accessed through the UK Data Service under the End User Licence. After obtaining and cleaning the data, I will replicate the ONS's original linear regression model as a baseline. Then, I will expand the set of features by adding more variables. To build predictive models, I will apply different machine learning techniques such as decision trees, random forests, and gradient boosting. These models will be trained using cross validation to reduce the risk of overfitting. I will also use SHAP to interpret the results and understand the positive or negative influence of each variable on life satisfaction.

Data availability and quality
The APS dataset is publicly available but requires UK-based access and registration. It is large and well-documented, but some features may have missing or inconsistent values. I will handle these through deleting them or data imputation and by carefully selecting variables with sufficient samples.

Technical challenges and risks
Some machine learning models may require longer training times or more computing resources. To manage this, I will start with simpler models and scale up as needed.

Resources
I will use Python and different supporting libraries like pandas, SHAP on my personal computer or university systems. Regular meetings with my supervisor will help ensure the project stays on track and any issues are resolved early.

Section B – Overall Programme (continued)

4. References

Please provide a list of references made in Sections B1, B2 and B3. The formatting of the references must comply with the Style Guide for Technical Reports and Academic Papers.

1. Clark, A.E. & Oswald, A.J., 2002. Well-being in panels. In: D. Kahneman, E. Diener & N. Schwarz, eds., *Well-being: The foundations of hedonic psychology*. New York: Russell Sage Foundation, pp.88–106.

2. De Neve, J.-E. & Oswald, A.J., 2012. Estimating the influence of life satisfaction and positive affect on later income using sibling fixed effects. *Proceedings of the National Academy of Sciences*, 109(49), pp.19953–19958. https://doi.org/10.1073/pnas.1211437109

3. Fernandes, M., Rocha, T. & Rodrigues, P.P., 2020. Machine learning approaches for predicting quality of life in cancer patients. *Health and Technology*, 10(1), pp.99–107. https://doi.org/10.1007/s12553-019-00345-0

4. Lundberg, S.M. & Lee, S.I., 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

5. Molnar, C., 2022. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed. Available at: https://christophm.github.io/interpretable-ml-book/ [Accessed 9 June 2025].

6. Nguyen, D. & Li, J., 2021. Predicting life satisfaction from socio-demographic variables using machine learning. *Information*, 12(6), p.246. https://doi.org/10.3390/info12060246

7. Office for National Statistics (ONS), 2019. *Economic associations with life satisfaction*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/personalandeconomicwellbeingintheuk/whatmattersmosttoourlifesatisfaction [Accessed 9 June 2025].

8. Office for National Statistics (ONS), 2023. *Personal well-being in the UK: May 2022 to May 2023*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/measuringnationalwellbeing/april2022tomarch2023 [Accessed 9 June 2025].

9. Powdthavee, N., 2008. Putting a price tag on friends, relatives, and neighbours: Using surveys of life satisfaction to value social relationships. *The Journal of Socio-Economics*, 37(4), pp.1459–1480. https://doi.org/10.1016/j.socec.2007.04.004

MSc Project Specification Form

Section C – Social, legal and ethical issues

Describe Social, legal and ethical issues that apply to your project (If your project requires a questionnaire/interview for conducting research and/or collecting data, you will need to apply for an ethical approval).

Does your project require ethical approval?

Ethical consideration
This project uses sensitive information collected, The analysis will aim to inform policy without reinforcing bias or stigma related to health, employment, or personal circumstances. Additionally, the use of explainable AI techniques like SHAP helps promote transparency and fairness in the interpretation of predictive models.

Legal considerations
Legal considerations include compliance with data protection regulations such as the UK GDPR and the Data Protection Act 2018. The project will ensure that all data is used strictly for academic purposes, with no attempts to identify individuals or relink data to specific respondents.

Social responsibility
As the project touches on well-being at a population level. Care will be taken in how findings are interpreted and communicated, especially in relation to vulnerable groups.

This project does not need ethical approve as long as it uses only publicly available, anonymized data. If any new data is collected in the future (e.g., through surveys), ethical approval would be sought accordingly

MSc Project Specification Form

Section D – Work plan

Task No	Task Describe each task to be undertaken and, where appropriate, indicate how it will be carried out. Please consider tasks between the submission of this specification until the submission of the dissertation in August.	Effort *(person weeks)*	Indicate deliverables and decision points (*e.g. data analysis completed, designs completed, software implemented, hardware procured, system configured,..*)
1	Background research and literature review-Review the ONS research methods and read articles about different predicting models.	2	Summary of article and model types
2	Understand and preprocess Data - Register with UK Data Service, download APS data, clean and preprocess dataset.	1	Download, clean data, handle missing data.
3	Implement the ONS's linear regression model ad baseline model	2	Recreate ONS model
4	Feature engineering and expansion – Understand and add meaningful features	2	Expanded feature set finalized and ready for modeling.
5	Develop Predictive Models – Build Lasso, Random Forest or Gradient Boosting models.	3	Train models with cross validation
6	Model evaluation and interpretation - Compare models to ONS baseline using R^2, RMSE, SHAP	2	Generate model comparison plots and SHAP plots
7	Discuss the results and findings	2	Summarize key findings
8	Dissertation writing	3	Full first draft completed
9	Review and Finalize	3	Dissertation complete

MSc Project Specification Form

Section E - Time Plan

For each task identified in Section D, shade the months during which you will be engaged and mark deliverables and decision points.

Task No	February	March	April	May	June	July	August	September
1					X			
2					X			
3						X		
4						X		
5							X	
6							X	
7							X	
8								X
9								X

MSc Project Specification Form

Section F – Costing

If your project requires any costing, give the justification and budget below. You need to consult your supervisor first before filling out this section and get approval from your supervisor. After the submission of this project specification form, the department will consider the budget to approve this or not. If your project does not require any costing, you may leave this section blank.

Justification of the budget	
Item description	Cost (£) *incl. VAT*
Total	

Appendix B

Peer Review Summary

CSMPR MSc Project
Peer Review Summary

Name of the reviewee:	Yiu Ting Wong
Title of the project being reviewed:	Effects of interactions and engagement with natural environment on self-rated health

Reviewer's comment:

Project Background:
The background is clearly stated, explaining why studying the natural environment and health is important and linking it to policy relevance. The PANS dataset is highlighted as the foundation of the project, and academic references are used to show both academic and applied importance. The research gap is also well defined: while descriptive reports exist, no statistical modelling has been applied. Some minor improvements could be made, such as adding a space before references (e.g., Jimenez et al., 2021) and clarifying why certain statistics are important. For example, the Natural England figures are interesting, but the decline from 94% to 92% over four years might not be statistically significant. Overall, the background is strong, relevant, and well-structured with a clear gap, only minor edits to phrasing and clarifications on the statistics would make it sharper.

Literature Review:
The literature review covers a wide range of studies, including international examples (China, Ireland, etc.), which demonstrates broad awareness. It not only reports findings but also engages with methodological approaches (e.g., logistic regression, XGBoost), showing technical understanding. Two research gaps are clearly identified: the lack of models using survey data on interaction and engagement with natural environments, and the overreliance on regression methods with limited exploration of explainable AI. Areas for improvement include integrating references more smoothly into continuous prose rather than listing them like bullet points, and discussing more limitations of the literature, such as potential biases from reliance on self-reported health data. Harvard-style citations are used but need corrections to spacing, grammar, and integration into sentences. With these refinements, the literature review would fully meet the criterion.

Research Question(s) / Problem Definition:
The research question is clearly formulated and directly aligned with the dataset, focusing on the associations between natural environment engagement and self-reported health. The supporting objectives are comprehensive, covering data preparation, exploratory analysis, baseline regression, advanced machine learning with explainable AI, and validation, which shows strong methodological awareness and an appropriate quantitative approach. However, the question could be sharpened to make it more hypothesis-driven. The objectives, while thorough, could also be streamlined into higher-level stages to improve readability and strengthen their direct link to the central question.

Development Approach:
The development approach is clearly described and follows a logical data science workflow, from exploratory analysis through data preparation, modelling, and evaluation. The methodology demonstrates strong awareness of technical requirements, including handling missingness, applying weighting to address imbalance, using regression and machine learning models, and incorporating explainable AI techniques such as SHAP and permutation feature importance. Risks related to data quality, self-reported bias, and computational constraints are acknowledged, with mitigation strategies proposed. One possible improvement would be to streamline the modelling plan, as it currently covers a wide range of approaches (regression, tree-based methods, deep learning). Narrowing the focus to the best performing or most relevant methods may make the project more feasible within the dissertation timeline. Overall, this section provides a well-structured and practical plan that reflects the intended computational methods and dataset.

Social, Legal, and Ethical Considerations:
The project demonstrates clear awareness of social, legal, and ethical considerations. Socially, the use of the dataset for non-commercial research is highlighted, along with risks of bias, overgeneralization, and potential stigmatization of groups. Mitigation strategies include representative sampling and careful communication of findings. Legally, compliance with the end-user license, UK GDPR, and the Data Protection Act is acknowledged, with strategies such as secure storage, non-sharing, and data deletion after project completion. This shows a solid understanding of data governance. Ethically, the project reflects good practice in confidentiality, transparency, minimizing confirmation bias, and considering potential adverse social impacts. Since no primary data collection is involved, the statement that ethical approval is not required is appropriate.

Project Objectives & Task/Time Planning:
The project objectives are clearly defined and well connected to the proposed tasks. Each task is described in detail, with logical sequencing from exploratory analysis through data preparation, modelling, evaluation, and finally dissertation writing and presentation. The plan shows good awareness of dependencies, such as pre-processing before modelling and tuning models before sensitivity analysis. Deliverables are identified at every stage, helping to track progress. The timeline also allows sufficient time for writing and reviewing the dissertation, which is crucial since mistakes in writing often reveal deeper issues that may require reworking parts of the analysis. While the plan is well structured, tasks like exploratory analysis and advanced modelling may require more flexibility than the fixed timelines suggest, given risks such as data quality or computational challenges. Additionally, preparation of the poster and presentation could start earlier to allow time for practice, given their importance to the overall MSc project.

MSc Project Specification Form

Version 03

Please complete this form in printed or typed text (Times New Roman size 11). **A copy of the approved document will have to be included as appendix in the actual dissertation** to help establish to which extend the project has been successful.

Section A

1. Project identification

Proposed dissertation title (maximum 15 words)	
Effects of interactions and engagement with natural environment on self-rated health	

2. Student details

Name (*in full*) Thomas Howell
e-mail address tghowell@student.reading.ac.uk

Term time address 32 Jessett Drive, Fleet, GU520XB
Contact telephone number 07717789951

3. Supervisor
Name and contact details of the staff member

Name Dr Ferran Espuny-Pujol
e-mail address

4. The supervisor *The person identified in Section A4 hereby approves the dissertation specification*

Name

Signature

Date

Fill in section A5 and A6, if the project has industrial input.

5. Company Partner
Name of organisation involved in the project

N/A

6. *Details of contact person in the organisation involved in the project*

Title *(e.g. Mr/Mrs/Dr)*	Name	
Address		
Tel. No.	Fax No.	e-mail address

MSc Project Specification Form

Section B – Overall Programme

1. Background and Literature review

Please describe in the space provided the background to the project and write a short literature review highlighting relevant developments on the topic of the dissertation

In an ever urbanising world, it is becoming increasingly important to understand the benefits that natural environments provide for personal health. Current research has shown evidence of associations between physical and mental health and exposure to natural environments**(Jimenez et al., 2021)**. To inform government policy Natural England and the Department for Environment, Food and Rural Affairs (DEFRA) conduct People and Nature Survey England (PANS) to gather information on public engagement and interaction with the natural environment.

Natural England publish an annual report of the key findings from the past year of the survey in which Year 1 94% **(Natural England, 2021)** of respondents who had visited a green space in the last 14 days agreed it had a positive impact on their physical health; this decreased to 93% in year 2 **(Natural England, 2022)** and 92% in years 3 **(Natural England, 2023)** and 4 **(Natural England, 2024)**.

Despite these descriptive insights no statistical modelling techniques have been applied to the set of surveyed factors gathered in the PANS dataset to explore the possibility of associations with participant's self-reported health outcomes. Researching associations between environmental factors and health beyond perceived benefits is crucial to understanding the true significance of relationships while accounting for confounding factors.

The project aims to address the gap by utilising statistical modelling techniques to determine the strength and direction of associations between people's use of and access to natural environments in relation to their self-reported health.

Literature Review

Research investigating the links between self-reported health and natural environments utilising different measures to ascertain the significance of associations.

- In China **(Huang et al., 2019)** used satellite imagery to extract percentages and proximity to greenspaces
- in Ireland **(O'Regan et al., 2021)** used Google Street View and satellite imagery to extract and quantify the Green View Index and Normalized Difference Vegetation Index for each participants location.
- Also in China **(Chen et al., 2023)** used air quality index as well as the NDVI for establishing relative greenness in the locality of participants
- In Bulgaria **(Dzhambov et al., 2023)** also investigated air pollution in addition to using NDVI as the primary environment exposure indicator with density of tree cover, number of trees and access to greenspaces within 300m used as secondary indicators of exposure.
- In England **(White et al., 2019)** used survey data to quantify the exposure to natural environment in the last 7 days in minutes.

All these studies established associations to self-reported health with the natural environment measures they used.

The methods used to analyse the strength and direction of the associations of the greenspace factors measured for and self-rated health.

- **(O'Regan et al., 2021)** used negative binomial regression to model for counts of health outcomes.
- **(Huang et al., 2019)** used logistic regression to categorise into good or bad health with the established greenspace factors a .

- **(Chen et al., 2023)** utilised XGBoost with **SHAPley** Additive exPlanations (SHAP) to establish ranking of importance of factors for good self-rated health.
- **(Dzhambov et al., 2023)** chose to use multivariate ordinal regression that included the natural environment measures being associated with poor self-rated health.
- **(White et al., 2019)** used weighted binomial regression to predict good health.

In summary recent approaches to modelling with self-rated health data typically combine data sources that establish natural environment factors about the locality of participants residence. Aside from **(White et al., 2019)** which uses surveyed time spent in natural environment, there is a notable gap in modelling using survey data that collects participant information on the types of interaction and engagement with specific natural environments visited. The PANS dataset contains an extensive collection of factors such as frequency of visits, time of visits, type of natural environment and activity at the location visited that provides the opportunity to research this gap in knowledge.

Furthermore with exception of **(Chen et al., 2023)** most studies limit modelling to directly interpretable regression models. Therefore, there is also an opportunity to explore the affects algorithms that capture more complexity can bring and utilise explainable AI techniques such as SHAP to understand the strength and nature of associations identified.

MSc Project Specification Form

Section B – Overall Programme (continued)

2. Research question, justification and objectives

Please describe in the space provided the research question to be answered, justify why the topic is important at the present time, and describe the specific objectives of research against which your achievements will be measured.

Research question:

Which natural environment engagement and interaction factors are associated with self-reported health?

Objectives:

- Determine, extract and define factors of environmental engagement and interaction to create a list of factors for modelling with self-reported health
- Analyse the distributions, patterns and correlations of features chosen for analysis as well as potential confounding factors and produce visualisations and a summary of the factor's descriptive statistics. Report on data quality to form a pre-processing strategy
- Pre-process and transform data to deal with missingness, imbalance and outliers by using appropriate data handling techniques to produce a clean dataset prepared with the chosen factors for modelling
- Create baseline regression models that predict self-reported health with interpretable results of direct associations between the chosen environmental factors and health
- Apply advanced ML algorithms to model natural environment factors that contribute to self-reported health outcomes utilising explainable AI to interpret the significance of factors contribution to health prediction to determine if significant complex relationships can be established
- Evaluate and validate associations by determining the extent of independence of any associations by performing stratified and sensitivity analysis to further establish the confidence of associations considering any assumptions or choices made in the prior stages of the project

MSc Project Specification Form

Section B – Overall Programme (continued)

3. Methodology

Please describe the methodology that you will use to achieve the objectives stated in Section B2.

You should also identify the data availability and quality, technical challenges, resource constraints, any other risks associate with the project and your plan to manage them.

The methodology to be adopted to complete the objective specified for the project will primarily follow a standard process for the completion of a data science research project. Below sets out the key stages the project will follow:

1. Exploratory Data Analysis

Create distribution charts to understand and compare how key features vary, for outcome measures and control factors including socioeconomic and demographic variables. Bivariate analysis will be conducted to explore interactions and relationships in the dataset's features. A table of summary statistics will be produced to give insight into average data values and variance. A report summarising the quality of the data will be produced that when combined with the prior analysis will enable a strategy to be formed to handle data quality.

2. Data Preparation

Determine and define each factor to be used for modelling as well as identify confounding factors to be kept for control of the experiments. Remove data not relevant to the analysis and deal with missingness by removing samples with no data for the outcome measure and impute, remove or otherwise handle records for other variables where appropriate. Process data to ensure imbalance is addressed through understanding and subsequent application of PANS sample weighting. Further over or under sampling may also be required to address further imbalance in the dataset that may be caused by removal of records as a part of handling other data quality issues. Scale and encode features as necessary dependent on the type of data a feature contains as well as the algorithm type's requirements.

3. Modelling

Using Python create baseline regression models to provide a baseline of interpretable models to understand basic relations between factors. Use tree-based models to research and understand if differences in significance of associations are found. Develop deep learning models to further research the complex non-linear relationships that may be present. Tune Hyper-parameters of all models created and cross validate to provide the most accurate results possible from the data. Leveraging both interpretable coefficients of baseline created along with explainable AI techniques including SHAP and PFI for models without direct interpretability, create statistics to rank and determine significance in the associations between variables modelled.

4. Results and Evaluation

Interpret the scores given to factors to draw conclusions as to the resulting association between factors modelled. Determine the significance of associations across confounding factors by performing stratified analysis to assess the relative difference in strength of association. Perform sensitivity analysis to understand how conclusions are affected by changes in data or model tuning. Create graphical and tabular representations to compare performance of models.

Data Availability and Risks

The data is available from the UK data service with access already having been granted based on the outline of the project, non-commercial academic purpose of the research and agreement to the end-user license.

Although PANS is an official statistic there is a risk of data quality issues especially given responses are self-reported which could be subject to recall or social desirability bias, to mitigate sample weights must be accounted for and sampling techniques used to adjust to make the conclusions drawn representative of the population. Also, missing data could lead to distortions in the strength of associations found and hence the conclusions made so this must be adjusted for prior to imbalance as removing records could skew the bias.

There is also a high risk of the delays in development affecting the time available for later stages especially given the short time for completion. To mitigate this risk work has been planned so that some tasks are beneficial but not an essential part of the process such as the number of models produced and sensitivity analysis can have their scope adjusted to ensure deadlines are met while not compromising the methodology to draw conclusions. Furthermore, computational constraints, although unlikely, could pose challenges to build models of the complexity required to adequately model and discover associations between features. This can be mitigated by making use of available departmental resource which is larger than what is available personally.

MSc Project Specification Form

Section B – Overall Programme (continued)

4. References

Please provide a list of references made in Sections B1, B2 and B3. The formatting of the references must comply with the Style Guide for Technical Reports and Academic Papers.

Chen, Y., Zhang, X., Grekousis, G., Huang, Y., Hua, F., Pan, Z., Liu, Y., 2023. Examining the importance of built and natural environment factors in predicting self-rated health in older adults: An extreme gradient boosting (XGBoost) approach. J. Clean. Prod. 413, 137432. https://doi.org/10.1016/j.jclepro.2023.137432

Dzhambov, A.M., Dimitrova, V., Germanova, N., Burov, A., Brezov, D., Hlebarov, I., Dimitrova, R., 2023. Joint associations and pathways from greenspace, traffic-related air pollution, and noise to poor self-rated general health: A population-based study in Sofia, Bulgaria. Environ. Res. 231, 116087. https://doi.org/10.1016/j.envres.2023.116087

Huang, B., Liu, Y., Feng, Z., Pearce, J.R., Wang, R., Zhang, Y., Chen, J., 2019. Residential exposure to natural outdoor environments and general health among older adults in Shanghai, China. Int. J. Equity Health 18, 178. https://doi.org/10.1186/s12939-019-1081-4

Jimenez, M.P., DeVille, N.V., Elliott, E.G., Schiff, J.E., Wilt, G.E., Hart, J.E., James, P., 2021. Associations between Nature Exposure and Health: A Review of the Evidence. Int. J. Environ. Res. Public. Health 18, 4790. https://doi.org/10.3390/ijerph18094790

Natural England, 2024. Adults' Year 4 Annual Report (April 2023 - March 2024) - GOV.UK.

Natural England, 2023. Adults' Year 3 Annual Report (April 2022 - March 2023) (Official Statistics).

Natural England, 2022. The People and Nature Survey for England: Year 2 Annual Report - Data and publications (April 2021 - March 2022) (Official Statistics) main findings.

Natural England, 2021. The People and Nature Survey for England: Data and publications from Adults survey year 1 (April 2020 - March 2021) (Official Statistics) main findings.

O'Regan, A.C., Hunter, R.F., Nyhan, M.M., 2021. "Biophilic Cities": Quantifying the Impact of Google Street View-Derived Greenspace Exposures on Socioeconomic Factors and Self-Reported Health. Environ. Sci. Technol. 55, 9063–9073. https://doi.org/10.1021/acs.est.1c01326

Parkes, O.L., 2024. End User Licence Agreement [WWW Document]. URL https://ukdataservice.ac.uk/app/uploads/cd137-enduserlicence.pdf

White, M.P., Alcock, I., Grellier, J., Wheeler, B.W., Hartig, T., Warber, S.L., Bone, A., Depledge, M.H., Fleming, L.E., 2019. Spending at least 120 minutes a week in nature is associated with good health and wellbeing. Sci. Rep. 9, 7730. https://doi.org/10.1038/s41598-019-44097-3

MSc Project Specification Form

Section C – Social, legal and ethical issues

Describe Social, legal and ethical issues that apply to your project (If your project requires a questionnaire/interview for conducting research and/or collecting data, you will need to apply for an ethical approval).

Does your project require ethical approval?

Social

The dataset used has been collected for non-commercial research, to maintain trust and transparency data must only be used for purposes set out when it is originally collected. To avoid bias results a representative sample must be used to avoid findings that could lead to decisions that adversely affect underrepresented groups as a result.
Implications of conclusions drawn can lead to stigmatisation of groups and to policy decisions that are bias or ineffective therefore findings should be communicated carefully with regard to potentially sensitive issues being addressed as well as avoiding overgeneralisation to reduce risk of biased or ineffective policies.

Legal

Access to the data is provided upon the agreement to an end-user license that sets out the conditions of use under which the data is provided**(Parkes, 2024)**. This includes obligations such as access must not be given to third parties or used with online tools such as LLMs without prior consent from the UKDA. Furthermore, access should only be from the UK and data can only be used for the specified non-commercial research project with copies of the data destroyed upon completion. These legal issues are mitigated by ensuring data is strictly processed for the specified research purpose, exclusively offline and within the UK.
Consideration must also be given to relevant UK laws governing personal data specifically UK GDPR and DPA. The end-user license implements some restrictions that enforce compliance with this legislation such as deletion upon completion of the research and no attempts to identify participants must be made. Furthermore, the usage and location limitations imposed will align with the agreement made with participants at data collection. To maintain this compliance the data must only be used for the research described when it is obtained and remain securely stored until it is destroyed as it is no longer required.

Ethical

Confidentiality must be kept for all participants to maintain trust with the population regarding how responses to potentially sensitive topics are handled. How the project benefits society more generally must be considered, and proactive action must be taken to minimise any negative consequences of the research and the impact of its findings.
The impact of confirmation bias must be mitigated by ensuring a methodology is maintained that tests hypotheses rigorously without consideration for expectations of results. Findings must be as transparent and explainable as possible especially given the black box nature of certain ML algorithms that may be used.

Ethical Approval

As no data collection is required for the project no ethical approval is necessary.

MSc Project Specification Form

Section D – Work plan

Task No	Task Describe each task to be undertaken and, where appropriate, indicate how it will be carried out. Please consider tasks between the submission of this specification until the submission of the dissertation in August.	Effort *(person weeks)*	Indicate deliverables and decision points (*e.g. data analysis completed, designs completed, software implemented, hardware procured, system configured,..*)
1	**Exploratory data analysis of distributions, descriptive statistics and bivariate analysis**	**7 days**	**Visualisations of distributions and relations between chosen features. Table summarising relevant descriptive statistics**
2	**Engineer and subsequently select natural environmental factors to be used for modelling**	**3 days**	**Decision on specific independent variables for analysis of association**
3	**Pre-process data by removing irrelevant data, addressing missingness in features and imbalance of variables**	**6 days**	**Clean dataset ready prepared for modelling**
4	**Utilise regression algorithms to create baseline models of associations of factors**	**5 days**	**Tuned directly interpretable regression models produced**
5	**Produce models that incorporate non-linearity to determine if significant associations are present with tree-based and deep learning methods. Determine associations by applying explainable AI techniques to establish feature importance**	**8 days**	**Tuned models incorporating potential non-linear relations produced and feature importance information gathered using explainable AI**
6	**Perform stratified and sensitivity analysis to determine effects of confounding factors as well as the robustness of any associations found**	**6 days**	**Factors importance identified relative to potential confounding factors**

MSc Project Specification Form

Section D – Work plan (continued)

Task No	Task Describe each task to be undertaken and, where appropriate, indicate how it will be carried out. Please consider tasks between the submission of this specification until the submission of the dissertation in August.	Effort *(person weeks)*	Indicate deliverables and decision points (*e.g. data analysis completed, designs completed, software implemented, hardware procured, system configured,..*)
7	**Prepare and write an introduction and literature review chapter for the MSc Project report. Detailing background, scope, aim, objectives, approach and a review of relevant literature that informs the project**	**3 days**	**Introduction and literature review chapters complete**
8	**Detail precisely the methodology used to complete each stage of the project. Along with relevant figures and tables that informed the process.**	**5 days**	**Methodology chapter complete**
9	**Explicitly record the results of the project using tables and/or diagrams to aid description of findings of the modelling conducted.**	**4 days**	**Results and discussion chapters complete**
10	**Summarise work conducted in reference to its aims and objectives and outline the research conclusions made. Reflect on the process of the project and detail future work on the subject.**	**1 day**	**Conclusions and Reflection Chapters complete**
11	**Create Poster communicating the project's process, results and conclusions to a diverse audience**	**2 days**	**Project Poster created**
12	**Create presentation slideshow to exhibit the findings of the report and practise along with ensuring the project code is formatted visually for demonstration**	**3 days**	**Report Slideshow and code notebook for demonstration**
13	**Review dissertation report to improve flow, reduce chance of errors in final submission and improve written quality.**	**5 days**	**Final Project Report for submission**

MSc Project Specification Form

Section E - Time Plan

For each task identified in Section D, shade the months during which you will be engaged and mark deliverables and decision points.

Task No	June	July	August	September
1	Exploratory Data Analysis			
2	Feature extraction and determination			
3		Data Pre-processing		
4		Baseline Regression modelling	Baseline Regression modelling	
5		Non-linear modelling	Non-linear modelling	
6			Sensitivity analysis	
7		Introduction/Lit Review	Introduction/Lit Review	
8		Methodology	Methodology	
9		Results	Results	
10			Conclusion	
11				Poster
12				Presentation
13				Final Report

MSc Project Specification Form

Section F – Costing

If your project requires any costing, give the justification and budget below. You need to **consult your supervisor first** before filling out this section and get **approval from your supervisor**. After the submission of this project specification form, the department will consider the budget to approve this or not. If your project does not require any costing, you may leave this section blank.

Justification of the budget
No Costs are anticipated for the research project to be completed

Item description	Cost (£) *incl. VAT*
Total	

Appendix C — Redundant features pruned by correlation (|r| ≥ 0.90)

Training-set Pearson correlation pruning — Discarded vs. kept representative

Discarded feature	Kept (correlated with)	|r|
CombinedAuthoritiescensus2021	CombinedAuthorities	1.0
LIVTOG	LIV12W	1.0
REFWKD	REFDTE	1.0
ITL321census2021	NUTS163	1.0
ITL321	NUTS163	1.0
THISQTR	QRTR	1.0
METHAL11	METHAL12	1.0
ILLSAT	ILLSUN	1.0
IN0792ER	INDE07R	1.0
DURUN	DURUN2	1.0
MF1664	SCHM12	0.999
TOTHRS	SUMHRS	0.999
ILLFRI	ILLSUN	0.999
ILLMON	ILLSUN	0.999
ILLTHU	ILLSUN	0.999
ILLWED	ILLSUN	0.998
ILLTUE	ILLSUN	0.998
NUTS162	NUTS163	0.998
ITL221	NUTS163	0.998
ITL221census2021	NUTS163	0.998
LIMITA	LIMITK	0.998
REDSTAT	REDCLOS	0.998
CONTUK	CAMEYR	0.998
FURN	TIED	0.998
IOUTCOME	CLAIMS14	0.998
ENROLL	CLAIMS14	0.997
QULNOW	CLAIMS14	0.997
MAINMS	METHSE1	0.997
PDWG10	SUPVIS	0.997
TTACHR	SUMHRS	0.997
HIQUL22D	LEVQUL22	0.997
IN0792SS	INDS07S	0.996

Appendix C — Redundant features pruned by correlation (|r| ≥ 0.90)

Training-set Pearson correlation pruning — Discarded vs. kept representative

Discarded feature	Kept (correlated with)	|r|
BENFTS	CLAIMS14	0.994
WRKLNG1	MAINRET	0.994
MAINME	METHMP01	0.993
CYMW	RELIGW	0.992
CYMR	RELIGW	0.992
CYMS	RELIGW	0.992
APPRCURR	CLAIMS14	0.991
CYMU	RELIGW	0.991
PAIDHRA	SUMHRS	0.991
METHAL01	MAINMA	0.99
APPR12	CLAIMS14	0.99
NSECMJ20	NSECM20	0.99
REDP1	REDCLOS	0.99
GOVTOF	NUTS163	0.99
ILLOFF	ILLSUN	0.99
JBTP101	WHYTMP6	0.989
ILLDAYS1	ILLSUN	0.989
USUHR	TOTUS2	0.989
REFWKY	FILEYEAR	0.989
ETH11S	RELIGS	0.987
CURED8	CLAIMS14	0.987
SOLOR	CONSEY	0.985
BACTHR	SUMHRS	0.984
FUTUR13	OOBEN	0.984
CRYOX7_EUL_Main	CRYOX7_EUL_Sub	0.983
REDIND	RDMPNO2	0.98
ILODEFR	INECAC05	0.98
DISEA	DISCURR20	0.98
REDOCC	RDMPNO2	0.979
METHM	LKSELA	0.978
ETHWHW	RELIGW	0.977
LKTIMA	LKSELA	0.977

Appendix C — Redundant features pruned by correlation (|r| ≥ 0.90)

Training-set Pearson correlation pruning — Discarded vs. kept representative

Discarded feature	Kept (correlated with)	|r|
LKSELC	LKTIMB	0.974
LLORD	TIED	0.972
JOBTYP2	INDS07S	0.972
USNET99	ERNFILT	0.972
SOLO2	SELF21	0.97
STAT2	INDS07S	0.97
Y2JOB	INDS07S	0.97
GOR9dcensus2021	NUTS163	0.968
GOR9d	NUTS163	0.968
NATOX7_EUL_Main	NATOX7_EUL_Sub	0.967
COURSE	ATTEND	0.967
UNEMBN1	UCREDIT	0.965
LOOK4	INECAC05	0.964
ETHWSC	RELIGS	0.963
USESLP	ERNFILT	0.963
ILLWK	INECAC05	0.963
RELBUS	INECAC05	0.962
TYEMPS	LKSELA	0.962
EVERWK	INECAC05	0.961
STATLR	REDYL13	0.961
OWNBUS	INECAC05	0.96
FTPT	INECAC05	0.96
PUBLICR	INECAC05	0.96
STATR	INECAC05	0.96
LKFTPA	LKSELA	0.959
MPNLR02	MANAGLR	0.959
SECJOB	INECAC05	0.959
DIFJOB	INECAC05	0.959
WNLEFT11	INECAC05	0.958
HOURLY	ERNFILT	0.958
WN2LFT11	INECAC05	0.957
AGE	AAGE	0.957

Appendix C — Redundant features pruned by correlation (|r| ≥ 0.90)

Training-set Pearson correlation pruning — Discarded vs. kept representative

Discarded feature	Kept (correlated with)	|r|
EVEROT	INECAC05	0.957
SECJMBR	INECAC05	0.956
SECTOR	INECAC05	0.956
FTPTW	INECAC05	0.954
PDWG102	MPNSR02	0.953
HOME	INECAC05	0.952
SUPVIS2	MPNSR02	0.952
MANAG2	MPNSR02	0.952
HITQUA15	CLAIMS14	0.951
INDE07M	INECAC05	0.948
IN0792EM	INECAC05	0.947
USGRS99	ERNFILT	0.946
EMPLEN	INECAC05	0.946
XDISDDA20	REDACT	0.945
HEALYL	HEALPB2001	0.945
SC20SMJ	SC20SMN	0.944
LKYT4	OOBEN	0.943
WAIT	OOBEN	0.943
JSATYP	UCREDIT	0.943
LIKEWK	OOBEN	0.937
LNGLST	DISCURR20	0.934
ACTWKDY2	ACTWKDY3	0.933
IN0792SM	INECAC05	0.932
PREFHR	LOOKM111	0.93
SC20MMJ	INECAC05	0.929
ED13WK	INECAC05	0.928
ETH11EW	ETHEWEUL	0.926
JOBTYP	RESTMR6	0.926
LEFTYR	INECAC05	0.925
TOTAC2	TOTUS2	0.925
LEFTM	REDYL13	0.924
ACTHR	TOTUS2	0.922

Appendix C — Redundant features pruned by correlation (|r| ≥ 0.90)

Training-set Pearson correlation pruning — Discarded vs. kept representative

Discarded feature	Kept (correlated with)	|r|
IN0792SL	REDYL13	0.919
TEN1	TIED	0.917
SC20LMJ	REDYL13	0.916
MANAGER	RESTMR6	0.914
TTUSHR	SUMHRS	0.914
CONMPY	RESTMR6	0.913
INDS07M	INECAC05	0.913
REFWKM	QRTR	0.912
BUSHR	SUMHRS	0.911
VARYHR	DIFFHR20	0.91
PAIDHRU	SUMHRS	0.91
ACTWKDY1	INECAC05	0.909
NETWK2	GRSSWK2	0.907
LKFTPC	LKTIMB	0.906
INDS07L	REDYL13	0.906
CTRY9D	RELIGE	0.905
MPNR02	INCNOW	0.903
TOTUS1	TOTAC1	0.902