Introduction

One of the most meaningful indicators of a country's progress is the wellbeing of its citizens. Wellbeing is not limited to physical health, it also includes 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. Unlike standard health measures, wellbeing reflects how people feel about their daily lives, their resilience in facing challenges, and their overall life satisfaction. This makes wellbeing an essential area of study for both healthcare and society.

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 possible before. Today, large amounts of information about individuals' lifestyles, habits, and health attributes can be collected and analysed. By applying machine learning techniques, it becomes possible to identify patterns within these data and predict someone's overall wellbeing. However, wellbeing is not a simple measurement, it is shaped by many interacting factors, making it a complex and challenging variable to model accurately. Most existing wellbeing studies rely on descriptive statistics or linear regression, and while these approaches provide broad insights, they struggle to capture non-linear interactions and often use large sets of predictors without addressing redundancy. Few studies systematically compare modern machine learning methods for wellbeing prediction while also emphasizing interpretability and feature efficiency. This project aims to address that gap.

My personal motivation for this project comes from a strong interest in the wellbeing of the UK population, particularly in the period after 2021. The COVID-19 pandemic, followed by wider social and economic challenges, has had a significant impact on people's health, lifestyles, and sense of security. Reports from the UK Office for National Statistics (ONS, 2018; ONS, 2019) and other organizations show clear changes in how people perceive their wellbeing, making this a particularly meaningful time to explore the issue.This study leverages the Annual Population Survey (APS, 2021–2023), which provides a unique opportunity to examine wellbeing during the post-pandemic recovery period. As an international student, I am also curious about how wellbeing is measured and understood within the UK context, and how data driven approaches can provide new insights into these changes.

The core aim of this project is 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.

The main objectives of this project are:

- **Data preparation** – cleaning, encoding, scaling, and balancing the dataset to ensure it is suitable for training machine learning models.
- **Model comparison** – training a wide range of models, including logistic regression, random forests, support vector machines, gradient boosting, k-nearest neighbours, and neural networks.
- **Hyperparameter tuning** – optimizing model settings systematically to achieve stronger and more reliable performance.
- **Evaluation** – measuring accuracy, F1-score, and ROC-AUC to assess model effectiveness and generalizability.

- **Interpretation** – discussing the results in the context of wellbeing research and highlighting what the outcomes might mean for future applications.

The motivation behind this study is twofold. First, it is academically valuable because wellbeing is a multidimensional and non-trivial target, making it a robust test case for comparing different algorithms. Second, it has practical significance: more accurate prediction of wellbeing could support healthcare providers, policymakers, and wellbeing-focused applications that offer individuals insights into their health. For policymakers, predictive models could inform targeted interventions for vulnerable groups, while for healthcare providers they could underpin personalized wellbeing support tools. In the UK context after 2021, such insights are particularly relevant, as they can help in understanding how health and social factors are shaping recovery and resilience. By combining data preparation, machine learning modelling, and critical evaluation, this project contributes to the growing field of computational wellbeing prediction. More importantly, it demonstrates how data science can be used to address real-world challenges that directly affect both individuals and society.

Social, Legal, Ethical Issues and Risk Management

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.

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. There is also a risk of stigmatizing groups identified as having “low wellbeing.”

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). 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 academic would require additional permissions. These measures together highlight the importance of strong legal compliance and transparent communication when scaling such predictive approaches.

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. Fairness is another key issue. Predictive analyses can inadvertently embed existing social and demographic biases, leading to underrepresentation of minority groups or systematic misclassification. 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. 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.

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. Model overfitting was another concern, addressed through cross validation, regularization, and systematic hyperparameter tuning. 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. 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 reviewing.

Literature Review

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 (WHO, 1948) well-known definition of health as "a state of complete physical, mental and social wellbeing, and not merely the absence of disease or infirmity." 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 lived 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. 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?” and 3) “Overall, how happy did you feel yesterday?” (ONS, 2018). Responses are recorded on a scale from 0 to 10, and in this project a composite wellbeing score was created by averaging the three answers. This score is then categorized into thresholds, with values of 0–4 labelled as “Low”, 5-6 labelled as “Medium”, 7-8 labelled as “High” and 9–10 as “Very High.” The strength of this approach lies in its simplicity and scalability, enabling comparisons across demographic and social groups at the national level. However, its brevity also limits its ability to capture the deeper, multi-dimensional aspects of wellbeing.

Although the ONS wellbeing questions are widely used in the UK, other instruments have been developed to provide more detailed measures. For example, the Warwick–Edinburgh Mental Well-Being Scale (WEMWBS) is a 14-item scale focusing on positive aspects of psychological functioning and has been validated in the UK population (Tennant et al., 2007). Similarly, the WHO-5 WellBeing Index offers a brief, 5-item self report tool that captures general mental wellbeing and has been extensively validated across cultures (Topp et al., 2015). For more holistic assessments, the Physical, Mental and Social Well-Being Scale (PMSW-21) combines physical health symptoms, emotional states, and social support, closely aligning with the WHO’s definition of wellbeing (Supranowicz & Paź, 2014). Another widely recognized tool is the SF-36 Health Survey, which measures health-related quality of life across eight domains, including physical functioning, social role limitations, and mental health, and is widely used in clinical research (Ware & Sherbourne, 1992). Compared with these measures, the ONS framework is shorter and easier to apply in large scale surveys, making it particularly useful for monitoring wellbeing at the national level. For the purpose of this project, the ONS framework is the most suitable, as it provides clear, standardized, and widely recognized measures of wellbeing in the UK. Nevertheless, it is acknowledged that wellbeing is a complex and multi-dimensional construct, and results must therefore be interpreted within the limits of the available data.

Wellbeing in the UK

Between 2020 and 2022, 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 (GOV.UK, 2020). These decreases reflect not just the immediate consequences of the health crisis but also prolonged societal anxiety and socioeconomic stress.

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, 2023). The Economics Observatory (2023) shows 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..Taken together, these findings underscore the timeliness of this project, which focuses on wellbeing during the post-pandemic recovery and amid ongoing economic pressures. Identifying the factors associated with very high wellbeing is particularly important, as both individual and societal resilience come under increasing pressure.

Determinants of wellbeing

Wellbeing is influenced by a complex interaction of social, economic, and health-related factors. Higher levels of trust, supportive relationships, and civic engagement have been found to significantly contribute to subjective wellbeing, making social connections and community engagement consistently strong predictors (Helliwell & Putnam, 2004). At the individual level, psychological factors such as optimism and resilience also play a key role, while maintaining good physical health and being free from chronic illness remain important determinants (Steptoe, Deaton, & Stone, 2015). Income and marital status also influence wellbeing; however, research indicates that the positive effects of income diminish after basic needs are met, emphasizing the greater significance of relational and psychological resources (Diener, Oishi, & Tay, 2018). In addition, housing stability and the quality of employment have been identified as important contextual drivers, shaping both financial security and social identity (ONS, 2019). This multidimensional nature of wellbeing reinforces the importance of adopting a modelling framework capable of capturing both structural conditions and individual resources when predicting wellbeing outcomes.

Machine learning in wellbeing research

Machine learning (ML) approaches have now been widely applied to wellbeing and health research, offering higher predictive accuracy and the ability to detect non-linear patterns compared to traditional models. Methods such as logistic regression and decision trees have historically been used for wellbeing prediction due to their simplicity and interpretability (Powdthavee et al., 2017), but with advances in computational methods, more sophisticated models are increasingly applied. For example, random forests improve predictive power by combining multiple decision trees and are effective at handling large sets of wellbeing predictors (Rodrigues et al., 2021). 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 (Lundberg et al., 2020). Neural network models, particularly multilayer perceptrons (MLPs), have also been applied to wellbeing and mental health studies, showing strong predictive performance when diverse behavioural, demographic, and health features are included, though their “black box” nature limits interpretability (Choi et al., 2020). 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).

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 MLP by identifying the relative importance of wellbeing predictors (Lundberg & Lee, 2017). Despite challenges such as algorithmic bias and interpretability trade-offs, the growing evidence suggests that tree-based ensembles and neural networks are among the most effective models for wellbeing research. For this project, these methods are particularly relevant because they combine predictive strength with interpretability, enabling not only the classification of wellbeing outcomes but also insights into the key factors driving very high wellbeing in the UK population.

Methodology

Data Collection

The dataset used in this study was obtained from the UK Data Service under Study Number 9291: Annual Population Survey (APS) Three-Year Pooled Dataset, January 2021 – December 2023. The APS is a large scale household survey that provides a wide range of socio-demographic, economic, and health related indicators across the UK. It was chosen for this project because it is nationally representative, collects standardized wellbeing questions used by the ONS, and provides a

sufficiently large sample size for robust machine learning analysis. The raw dataset contained 341,465 respondents and 459 attributes. Special missing codes were embedded within the dataset, where values such as "−8" (No answer) and "−9" (Does not apply) indicated non response. These codes required systematic cleaning before analysis.

Exploratory Data Analysis

An initial missing value analysis revealed that three variables — GRSSWK, NETWK, and RELIG11 — contained more than 5% missing values and were therefore excluded from the dataset. All other variables exhibited less than 1.6% missingness and were retained.

The main outcome of interest, subjective wellbeing, was measured using three survey items: life satisfaction (SATIS), sense of worth (WORTH), and happiness (HAPPY). After excluding "−8" and "−9" codes, the column means were calculated as 7.55, 7.86, and 7.51 respectively. The missing responses were then imputed with these column means, ensuring comparability across respondents and enabling the calculation of a composite Wellbeing score, defined as the row wise average of the three indicators.

To better understand the distribution of this score, Figure 1 shows the histogram of raw wellbeing values across respondents. The distribution is right skewed, with the majority of individuals reporting scores above 6. Notably, approximately 50% of the dataset scored above 8, which provided a natural threshold for binary conversion. Based on this observation, the composite Wellbeing score was further transformed into a new target variable, Wellbeing_category, where scores below 8 were labelled 0 ("Low–High") and scores of 8 or above were labelled 1 ("Very High"). The resulting class balance is shown in Figure 2, where the two categories are nearly evenly distributed (≈49% vs. 51%). This balanced split ensured that the classification task was not biased toward a majority class. This approach aligns with ONS practice of using the 0–10 wellbeing scale, while simplifying it into a binary outcome suitable for predictive modelling.

Figure 1. Distribution of Raw Wellbeing Scores

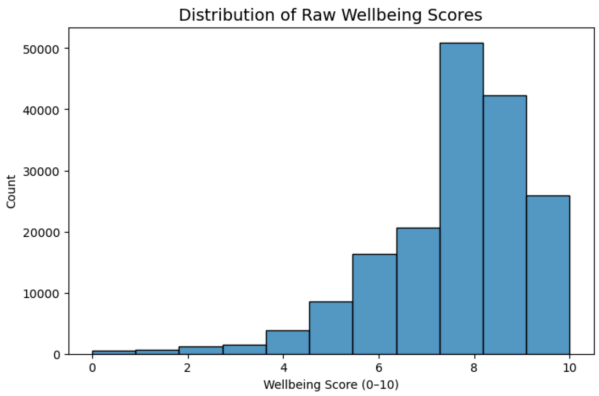
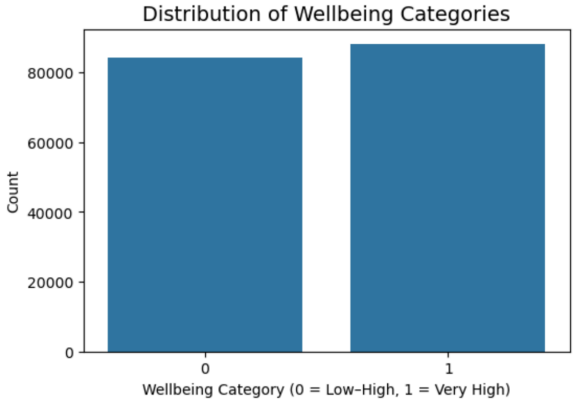


Figure 2. Distribution of Wellbeing Categories



Feature engineering and selection

Health condition variables HEALPB2001–HEALPB2010, which originally encoded conditions as numeric codes. These were recoded into 18 binary indicators (e.g., HEALTH_MentalAnxiety, HEALTH_HeartOrCirculation, HEALTH_Autism), where 1 indicated the presence of the condition. Binary encoding was chosen over ordinal representations to reduce noise, avoid false ordering, and improve interpretability. Finally 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.

To further reduce dimensionality and improve interpretability, feature selection was carried out using Normalized Mutual Information (NMI) between each attribute and the target. The strongest predictor was ANXIOUS (NMI = 0.06), followed by several health and employment related features. To avoid redundancy, attributes highly correlated with already selected features (correlation ≥ 0.9) were excluded. A Random Forest classifier was then used to evaluate performance as the number of top ranked features increased. Results showed that models trained on as few as 5–8 features achieved almost the same performance (AUC ≈ 0.75, Accuracy ≈ 0.69, F1 $\approx$ 0.69) as models using over 170 features. Based on this analysis, the top 7 non redundant features were selected for model training, as performance plateaued beyond this point.

Algorithms and Computational Processing

Following data cleaning and transformation, the dataset was prepared for predictive modelling. Categorical features were encoded using label encoding, binary variables were preserved in 0/1 format, and continuous variables were standardized for algorithms sensitive to feature scales. Specifically, continuous predictors were standardized using Z-score scaling for the MLP Neural Network, as neural networks are highly sensitive to feature magnitude. In contrast, tree based models (Random Forest, XGBoost) and linear classifiers were trained directly on raw features to retain interpretability and avoid unnecessary transformations.

The dataset was partitioned into training (70%), validation (15%), and independent test (15%) subsets using stratified sampling, which ensured that the balanced class distribution ($\approx$49% vs. 51%) was preserved across all splits. This design ensured fairness and prevents data leakage in model evaluation by using the training/validation sets for feature selection, model development, and hyperparameter tuning, while strictly reserving the independent test set for unbiased performance estimation.

A range of machine learning algorithms were evaluated, spanning linear models (Logistic Regression, Ridge Classifier, Linear SVM), distance based approaches (K-Nearest Neighbours), ensemble methods (Random Forest, XGBoost), and neural architectures (Multi-Layer Perceptron). This algorithmic breadth allowed comparison between interpretable baselines and more complex models with higher predictive capacity. Hyperparameter optimization was conducted using GridSearchCV with stratified 3-fold cross validation, a choice balancing computational efficiency with reliable variance estimation given the dataset's size. Model performance was evaluated using Accuracy, Weighted F1-Score, and ROC-AUC, with F1 and AUC prioritized due to their robustness under mild class imbalance. The Random Forest and MLP Neural Network consistently emerged as the strongest performers, and these were further analyzed using SHAP (SHapley Additive Explanations) to identify the most influential predictors of wellbeing and to improve interpretability.

Implementation

Setup

All experiments were conducted in Python 3.12.4 (Anaconda distribution), using Jupyter Lab 4.0.11 as the interactive environment. The main libraries included: pandas, numpy (data handling), scikit-learn (preprocessing, feature selection, model training, evaluation), xgboost (gradient boosting), shap (explainability), and matplotlib (visualization). Categorical features were encoded using label encoding, binary variables were preserved as 0/1, and continuous predictors were standardized with Z-score scaling when required (e.g., MLP Neural Network). Tree-based models and linear classifiers were trained directly on raw features since they are scale-invariant.

The dataset was split into training (70%), validation (15%), and test (15%) subsets using stratified sampling, which preserved class balance. Training and validation sets were used for feature selection, model development, and hyperparameter tuning, while the test set was reserved strictly for final evaluation.

Model Training and Hyperparameter Tuning
Seven algorithms were evaluated: Logistic Regression, Ridge Classifier, Linear SVM, K-Nearest Neighbours, Random Forest, XGBoost, and a Multi-Layer Perceptron (MLP Neural Network). This range was chosen to cover baseline linear models, distance-based approaches, tree-based ensembles, and neural networks, allowing both interpretability and predictive strength to be assessed.

Each model was fine-tuned using GridSearchCV with stratified 3-fold cross-validation on the training set. Hyperparameter search spaces were defined individually (e.g., tree depth and learning rate for XGBoost, regularization strength for linear models, neighbours for KNN, hidden layers and learning rate for MLP). Validation performance was then used to select the optimal hyperparameter, based on Accuracy, Weighted F1-score, and ROC-AUC.

After tuning, the best version of each model was retrained on the combined training and validation sets to maximize data availability. Final evaluations were performed on the independent test set, and results were summarized in comparative tables and diagnostic plots.

Evaluation and Interpretability
Model performance was assessed using Accuracy, Weighted F1-Score, and ROC-AUC, ensuring that both overall predictive ability and class balance were captured. Confusion matrices and classification reports were produced for each tuned model, providing insight into precision, recall, and misclassification patterns.

Beyond raw predictive performance, interpretability was addressed using SHAP (SHapley Additive Explanations). Two models were selected for analysis: XGBoost, explained using TreeExplainer, and the MLP Neural Network, explained using KernelExplainer with sampled subsets for computational efficiency. The use of both TreeExplainer (for ensemble models) and KernelExplainer (for neural networks) enabled consistent interpretability across different algorithmic families. Global importance was illustrated through summary bar plots, while bee swarm plots showed how individual feature values influenced predictions. This dual perspective—global feature ranking and local decision pathways—provided explanatory depth, highlighting predictors such as ANXIOUS and LIMACT as dominant drivers of wellbeing classification. This step directly links back to wellbeing literature by showing how psychological and activity-related measures align with known determinants of subjective wellbeing.

Results

The hyperparameter tuning stage demonstrated that XGBoost and Random Forest achieved the highest validation performance, both with ROC-AUC values of approximately 0.755, closely followed by the MLP Neural Network (ROC-AUC = 0.754). Linear models such as Ridge Classifier, Logistic Regression, and Linear SVM showed slightly lower performance (ROC-AUC ≈ 0.746), while KNN's performance is a little behind with a ROC-AUC of 0.709. These results indicate that ensemble tree based methods and neural networks were more effective in capturing nonlinear relationships in wellbeing data (Table 1).

Table 1. Validation performance across 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

On the independent test set, performance patterns were consistent with validation. XGBoost achieved the strongest results overall (Accuracy = 0.698, F1 = 0.698, ROC-AUC = 0.758), closely followed by the MLP Neural Network (ROC-AUC = 0.757) and Random Forest (ROC-AUC = 0.757). Ridge Classifier, Logistic Regression, and Linear SVM scored slightly lower (ROC-AUC ≈ 0.747), while KNN again underperformed (ROC-AUC = 0.712). These findings confirm that nonlinear models outperformed linear approaches in predicting wellbeing. This consistency reinforces that non-linear models consistently outperformed linear baselines in predicting wellbeing (Table 2).

Table 2. Final test set performance across models.

	Accuracy	F1–score	ROC–AUC
XGBoost	0.6978	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

Confusion matrices provided further insights into classification behaviour. For XGBoost, 8,977 “Low-High wellbeing” cases and 9,061 “Very high wellbeing” cases were correctly classified, with 3,671 and 4,141 misclassifications respectively (Figure 3). The MLP Neural Network produced a highly comparable confusion matrix, correctly identifying 8,736 “Low-High wellbeing” cases and 9,299 “ Very high wellbeing” cases, while misclassifying 3,912 and 3,903 instances (Figure 4). These results highlight the balanced nature of errors across classes, with overall accuracies close to 70%.

Figure 3. Confusion matrix for XGBoost.

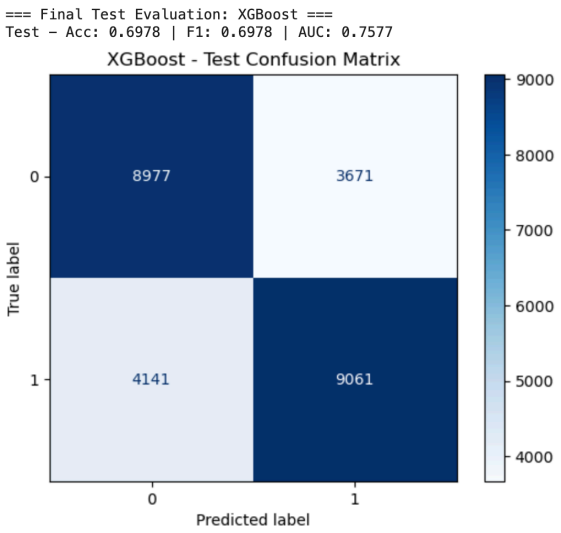
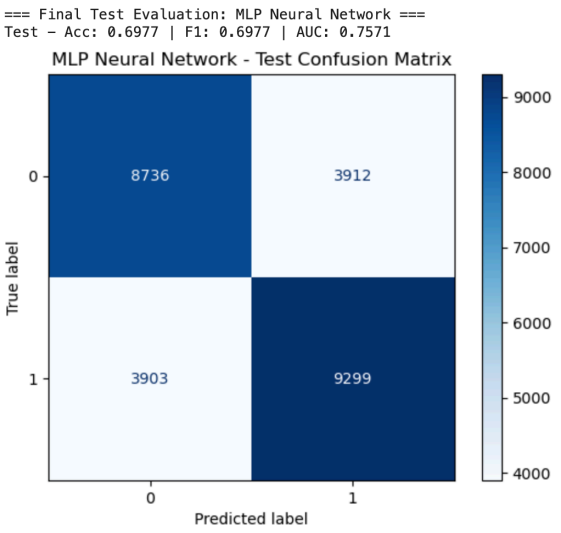


Figure 4. Confusion matrix for MLP Neural Network.



To enhance interpretability, SHAP (SHapley Additive Explanations) was applied to XGBoost and the MLP Neural Network. For XGBoost, ANXIOUS emerged as the most influential feature (mean |SHAP| ≈ 0.80), far exceeding contributions from MARCHK (0.20), REDACT (0.17), MARDY6 (0.12), and physical limitation measures such as LIMACT and LIMITK. A similar pattern was observed in the MLP model, where ANXIOUS again dominated (mean |SHAP| ≈ 0.41), followed by REDACT, LIMACT, and LIMITK. This consistency across models underscores the critical role of psychological distress and physical health limitations in shaping wellbeing outcomes (Figures 5).

Figure 5. SHAP feature importance bar plots for XGBoost and MLP Neural Network

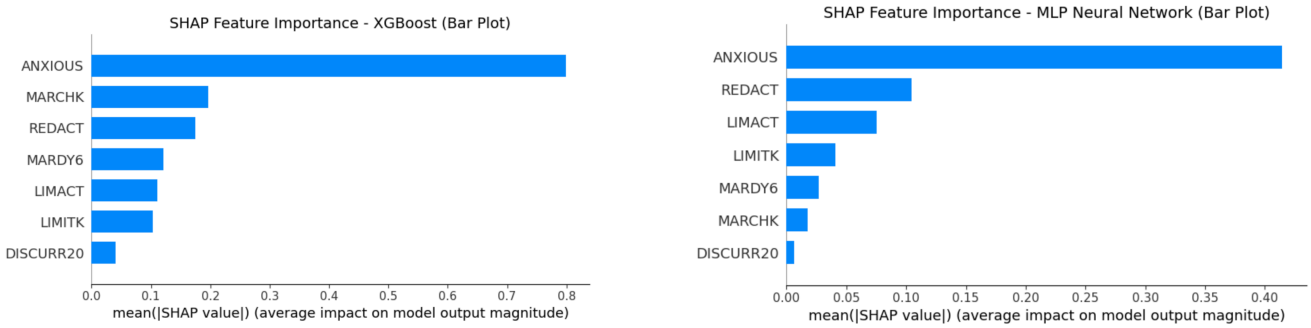
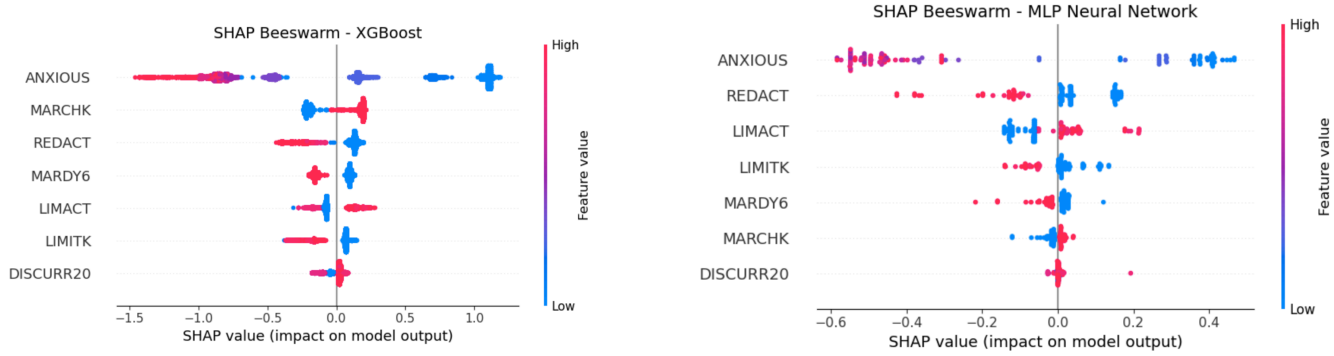


Figure 6. SHAP bee swarm plots for XGBoost and MLP Neural Network



The bee swarm plots further illustrated how individual feature values influenced predictions. Higher anxiety scores were strongly associated with lower predicted wellbeing, while fewer reported limitations in daily activity (e.g., LIMACT, LIMITK) were associated with higher wellbeing predictions. These findings provide actionable insights by confirming the importance of both psychological and physical health factors in wellbeing classification.

Discussion

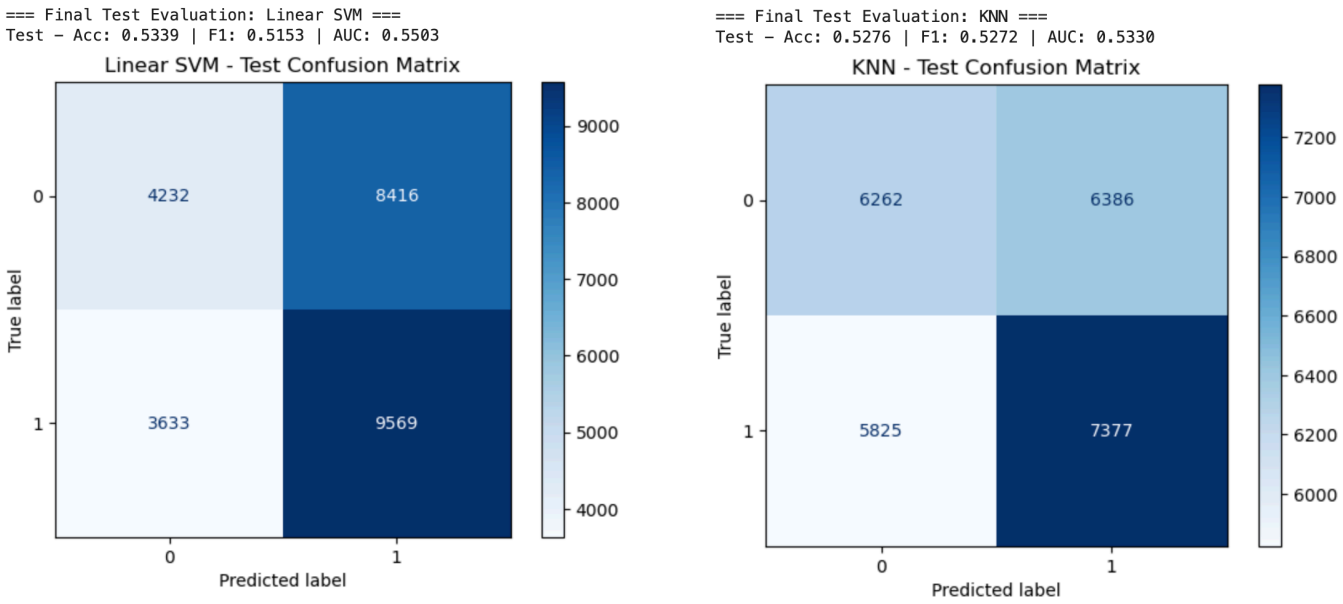
The results demonstrate that the modelling framework successfully captured key determinants of wellbeing, while maintaining methodological strictness. The predictive performance across models was consistent, with XGBoost, Random Forest, and the MLP neural network achieving the strongest results (Accuracy ≈ 0.70, AUC ≈ 0.75). These outcomes validate the modelling pipeline and align with findings in the wellbeing literature, which consistently identify mental health for example anxiety levels, physical limitations, and social circumstances as the strongest correlates of self reported wellbeing. Importantly, the fact that predictive performance plateaued after the inclusion of approximately seven features demonstrates that the most influential determinants of wellbeing can be distilled into a small subset of variables without significant loss in accuracy. This supports the robustness of the feature selection strategy, which combined Normalized Mutual Information with correlation pruning and model based validation.

Table 3. Final test set performance across models (178 features).

	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

To further test robustness, an alternative setup using top 178 features instead of top 7 was evaluated. As shown in Table 3, the top performing models (XGBoost, Random Forest, and MLP) achieved almost identical outcomes across both setups, with a slight improvement in AUC (≈0.76–0.77 vs. ≈0.75). 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. In 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 (≈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 7 (Confusion Matrices, 178 features), which show that high dimensional noise disproportionately increased misclassifications for these weaker algorithms.

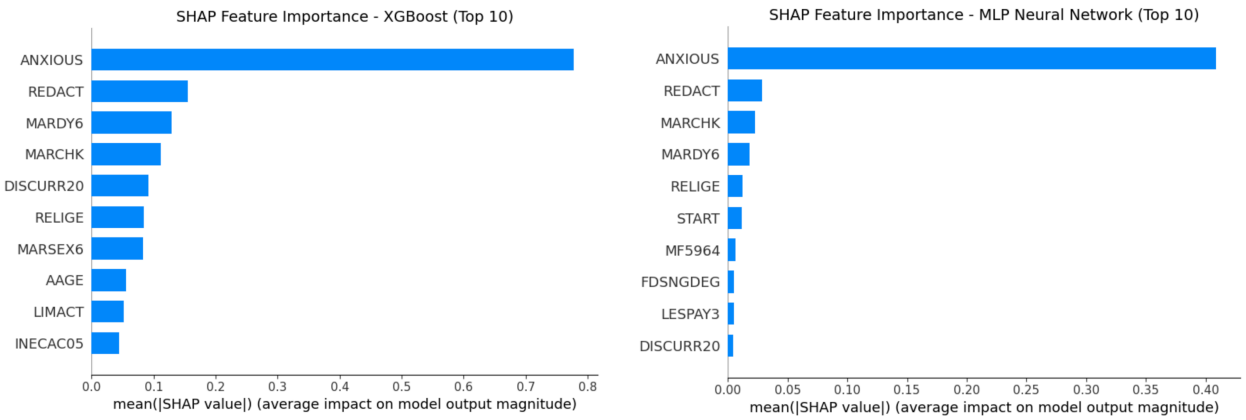
Figure 7. Confusion matrix for SVM and KNN (178 features)



Importantly, SHAP analyses (Figure 8) confirmed that the same core features — ANXIOUS, LIMACT, and REDACT dominated predictions in both the 7 feature and 178 feature setups. This reinforces that the additional 171 attributes contributed little new signal. Consequently, the 7 feature framework represents the most efficient and interpretable solution, retaining nearly identical predictive strength while avoiding the instability and complexity of high dimensional models.

Additional robustness trials further validated this conclusion. A binary split at 4 produced superficially high accuracy (≈0.97) but was driven entirely by severe class imbalance, with models effectively defaulting to the majority class. Similarly, a four-class categorization aligned with ONS achieved moderate accuracy (≈0.59) and AUC (≈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 (≈49% vs. 51%), stable and consistent performance across algorithms, and the clearest trade-off between predictive power and interpretability.Taken together, these robustness checks confirm that the binary split at 8, paired with a compact 7-feature set, offers the most reliable and interpretable framework for wellbeing prediction.

Figure 8. SHAP feature importance bar plots for XGBoost and MLP Neural Network (178 features)



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 in data, the predictive performance remained highly consistent with the main pipeline (XGBoost AUC = 0.754, Accuracy = 0.697; MLP AUC = 0.754, Accuracy = 0.699; Random Forest AUC = 0.753, Accuracy = 0.697). Interestingly, results were slightly lower overall, suggesting that "−8" and "−9" codes may themselves carry weak predictive signal, likely reflecting patterns in survey response behaviour or contextual applicability. This reinforces that aggressive exclusion of such codes does not improve model performance and that the main pipeline, which retains them, provides a more balanced and robust treatment.

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 understated 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. Similarly, the MLP Neural Network leverages hidden layers and non linear activation functions, enabling it to approximate complex mappings between predictors and wellbeing categories. In 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 highlight their suitability for modelling multifactorial constructs like wellbeing, where interactions between predictors are critical.

From a verification perspective, the project demonstrated to best practices in data science. Data splitting into train, validation, and independent test sets ensured unbiased performance estimation and data leakage. Hyperparameter tuning was performed systematically using stratified cross validation, and scaling was applied selectively to models requiring it for example MLP. Furthermore, model interpretability was addressed through SHAP analysis, which highlighted ANXIOUS, REDACT, and LIMACT as the most influential predictors, findings that resonate strongly with prior wellbeing research. These results justify the project's achievements, demonstrating that it not only replicated known associations but also provided quantitative evidence of their relative predictive power.

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. 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.

Originality of contribution

his project makes an original contribution to wellbeing research by combining advanced machine learning methods with the field's traditional reliance on regression analysis. For instance, the Office for National Statistics (ONS, 2021) and other large-scale wellbeing studies often employ regression with many socio-economic and health features, producing interpretable but limited predictive models (Diener et al., 2018; Helliwell & Putnam, 2004). In contrast, this study reframed wellbeing prediction as a binary classification problem and systematically reduced the predictor set from 471 to just seven variables using Normalized Mutual Information (NMI) and correlation pruning. Importantly, this reduction preserved predictive performance (AUC ≈ 0.75), demonstrating that wellbeing outcomes can be modelled effectively with a compact, non-redundant set of features. This is significant because feature-efficient modelling is uncommon in wellbeing analytics, where large predictor sets are usually retained without addressing redundancy (Oparina et al., 2022; Rodrigues et al., 2021). By showing that seven features are sufficient, the study introduces a novel pathway toward leaner wellbeing surveys that reduce respondent burden while maintaining predictive power with clear methodological and practical implications.

The second novel element is the integration of predictive strength with interpretability. Rather than relying on simpler but less accurate models, this study applied XGBoost and MLP neural networks, methods often criticized as "black boxes" and then used SHAP explainability tools (TreeExplainer and KernelExplainer) to uncover the role of each feature in both global and individual-level predictions (Lundberg & Lee, 2017; Lundberg et al., 2020). This dual focus bridges a persistent gap in wellbeing research, where interpretability has historically been prioritized at the expense of predictive accuracy (Steptoe et al., 2015; ONS, 2018). By demonstrating that complex models can be made transparent, this work advances the methodological toolkit for wellbeing science.

The third contribution lies in the systematic robustness testing. By comparing alternative modelling setups — using 178 features, a threshold at 4, and a four-class categorization, the study shows that its main framework (binary split at 8 with 7 features) achieves the most stable and interpretable balance. Such explicit testing of design choices is rarely carried out in wellbeing studies and enhances both the credibility and the originality of the modelling framework.

The results are therefore both quantitatively robust and practically useful: policymakers can identify key predictors such as anxiety and physical limitations, while researchers gain confidence in the interpretability of advanced models. Taken together, these contributions establish a foundation for more efficient, interpretable, and policy relevant wellbeing modelling, while providing a roadmap for future research in survey optimization and predictive public health.

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 encoding and preprocessing approaches. 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 SMOTE for class balancing and SHAP for interpretability 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, where even after applying SMOTE the minority classes in alternative formulations remained difficult to predict, with near zero recall in some cases. 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 precision. These challenges taught me the importance of flexibility, critical thinking, and resource management in data science projects.

From a personal perspective, this project fostered substantial growth. I learned the value of being willing to start over and over again, rather than being constrained by existing work or assumptions. I developed the habit of not only checking performance metrics but also examining model behaviour directly through confusion matrices and interpretability tools, which revealed weaknesses hidden behind seemingly strong scores. I also improved in project management, recognizing that building a strong coding pipeline from data preparation through modelling, evaluation, and interpretation is as important as the models themselves. Perhaps most importantly, I gained confidence in connecting data science methods to real world social and health challenges, which strengthened my motivation to continue in this field.

Looking ahead, there are clear directions for extending this work. One avenue would be to experiment with deeper and more complex models, such as advanced neural architectures, which may capture subtle non-linear interactions that traditional methods miss. Another would be to explore additional or alternative predictors, potentially integrating longitudinal data or richer behavioural variables. Finally, future research could investigate more sophisticated strategies for handling imbalance and bias, improving model reliability in minority groups. Overall, this project has not only strengthened my technical and academic skillset but also provided a foundation for future exploration of how machine learning can meaningfully contribute to wellbeing research.

References

Choi, J., Kim, H. & Lee, S., 2020. Machine learning approaches to mental health prediction using large-scale health and survey data. *BMC Medical Informatics and Decision Making*, 20(Suppl 11), p.302. https://doi.org/10.1186/s12911-020-01361-0

Diener, E., Oishi, S. & Tay, L., 2018. Advances in subjective well-being research. *Nature Human Behaviour*, 2(4), pp.253–260. https://doi.org/10.1038/s41562-018-0307-6

Economics Observatory, 2023. How is the cost of living crisis affecting public health? Available at: https://www.economicsobservatory.com/how-is-the-cost-of-living-crisis-affecting-public-health [Accessed 18 August 2025].

GOV.UK, 2020. *Wellbeing and loneliness – Community Life COVID-19 Re-contact Survey 2020 Main Report*. Available at: https://www.gov.uk/government/statistics/community-life-covid-19-re-contact-survey-2020-main-report [Accessed 18 August 2025].

Helliwell, J.F. & Putnam, R.D., 2004. The social context of well-being. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359(1449), pp.1435–1446. https://doi.org/10.1098/rstb.2004.1522

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

King's College London, 2023. *Cost-of-living crisis is worsening the mental health of most vulnerable*. Available at: https://www.kcl.ac.uk/news/cost-of-living-crisis-is-worsening-the-mental-health-of-most-vulnerable [Accessed 18 August 2025].

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

Lundberg, S.M., Erion, G.G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., … & Lee, S.I., 2020. From local explanations to global understanding with explainable AI for tree-based models. *Nature Machine Intelligence*, 2(1), pp.252–259. https://doi.org/10.1038/s42256-019-0138-9

Mind, 2023. *Mental health of half of adults in England and Wales negatively affected by cost-of-living crisis*. Available at: https://www.mind.org.uk/news-campaigns/campaigns/benefits/cost-of-living-crisis/ [Accessed 18 August 2025].

Office for National Statistics (ONS), 2018. *Measuring national well-being: Personal well-being in the UK*. Available at: https://www.ons.gov.uk [Accessed 18 August 2025].

Office for National Statistics (ONS), 2019. *Personal and economic wellbeing in the UK: May 2019*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/personaland economicwellbeingintheuk/may2019 [Accessed 18 August 2025].

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

Powdthavee, N., Lekfuangfu, W.N. & Wooden, M., 2017. What's the good of education on our overall quality of life? A simultaneous equation model of education and life satisfaction for Australia. *Journal of Behavioral and Experimental Economics*, 63, pp.55–63. https://doi.org/10.1016/j.socec.2016.04.001

Rodrigues, T., Filgueiras, A. & Alencar, M., 2021. Using random forests to predict quality of life and well-being based on health survey data. *International Journal of Medical Informatics*, 149, p.104429. https://doi.org/10.1016/j.ijmedinf.2021.104429

Steptoe, A., Deaton, A. & Stone, A.A., 2015. Subjective wellbeing, health, and ageing. *The Lancet*, 385(9968), pp.640–648. https://doi.org/10.1016/S0140-6736(13)61489-0

Supranowicz, P. & Paź, M., 2014. Holistic measurement of well-being: Psychometric properties of the Physical, Mental and Social Well-Being Scale (PMSW-21) for adults. *Roczniki Państwowego Zakładu Higieny*, 65(3), pp.251–258.

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., … & Stewart-Brown, S., 2007. The Warwick–Edinburgh Mental Well-being Scale (WEMWBS): Development and validation. *Health and Quality of Life Outcomes*, 5, p.63.

Topp, C.W., Østergaard, S.D., Søndergaard, S. & Bech, P., 2015. The WHO-5 Well-being Index: A systematic review of the literature. *Psychotherapy and Psychosomatics*, 84(3), pp.167–176.

Ware, J.E. & Sherbourne, C.D., 1992. The MOS 36-item short-form health survey (SF-36). *Medical Care*, 30(6), pp.473–483.