# scientific reports

### OPEN



## <sup>N</sup> Using machine learning to explore the efficacy of administrative variables in prediction of subjective-wellbeing outcomes in New Zealand

Anantha Narayanan<sup>1</sup>, Tom Stewart<sup>1⊠</sup>, Scott Duncan<sup>1</sup> & Gail Pacheco<sup>2</sup>

The growing acknowledgment of population wellbeing as a key indicator of societal prosperity has propelled governments worldwide to devise policies aimed at improving their citizens' overall wellbeing. In New Zealand, the General Social Survey provides wellbeing metrics for a representative subset of the population (~10,000 individuals). However, this sample size only provides a surface-level understanding of the country's wellbeing landscape, limiting our ability to comprehensively assess the impacts of governmental policies, particularly on smaller subgroups who may be of high policy interest. To overcome this challenge, comprehensive population-level wellbeing data is imperative. Leveraging New Zealand's Integrated Data Infrastructure, this study developed and validated the efficacy of three predictive models—Stepwise Linear Regression, Elastic Net Regression, and Random Forest-for predicting subjective wellbeing outcomes (life satisfaction, life worthwhileness, family wellbeing, and mental wellbeing) using census-level administrative variables as predictors. Our results demonstrated the Random Forest model's effectiveness in predicting subjective wellbeing, reflected in low RMSE values (~1.5). Nonetheless, the models exhibited low R<sup>2</sup> values, suggesting limited explanatory capacity for the nuanced variability in outcome variables. While achieving reasonable predictive accuracy, our findings underscore the necessity for further model refinements to enhance the prediction of subjective wellbeing outcomes.

Keywords Subjective wellbeing, Machine learning, Predictive models, Administrative data, Census

The significance of population wellbeing is gaining widespread recognition globally, prompting governments to broaden their evaluative criteria beyond the traditional measure of GDP (Gross Domestic Product) to assess the overall success of their population<sup>1</sup>. While GDP and productivity measures continue to be central for policymaking, there is an emerging shift towards a more comprehensive approach that includes the assessment of wellbeing. Initiatives like the Wellbeing Economy Governments partnership (WEGo) exemplify this shift, where national and regional governments collaboratively advance the concept of Wellbeing Economies<sup>2</sup>. Despite sustained economic growth, New Zealand faces pressing challenges such as high rates of child poverty, homelessness, and suicide. In response, the government introduced its inaugural 'wellbeing budget' in 2019<sup>3</sup>, signifying a renewed commitment to prioritising people's wellbeing alongside economic growth.

Understanding wellbeing presents challenges due to the evolving nature and diverse perspectives around its meaning. Initially, wellbeing was often perceived as positive human functioning, referred to as "eudaimonia," encompassing aspects such as self-actualisation and autonomy<sup>4</sup>. Other researchers have integrated eudaemonic and hedonic components, combining aspects of functioning and emotions<sup>5</sup>. For example, Diener's tripartite model identified cognitive, positive affect, and negative affect components<sup>6</sup>, while Seligman's PERMA model introduced positive emotion, engagement, relationships, meaning, and accomplishment as key dimensions<sup>7</sup>. Thompson et al.'s dynamic model of 'flourishing' further highlights the interplay between positive feelings, effective functioning, external conditions, and personal resources<sup>8</sup>. This comprehensive perspective suggests that 'flourishing' or elevated wellbeing emerges from the interplay of positive emotions and effective functioning

<sup>1</sup>School of Sport and Recreation, Auckland University of Technology, Private Bag 92006, Auckland 1142, New Zealand. <sup>2</sup>Faculty of Business, Economics and Law, Auckland University of Technology, Auckland, New Zealand. <sup>22</sup>email: tom.stewart@aut.ac.nz

within an individual's unique circumstances and available resources. Thus, a 'flourishing' nation indicates elevated wellbeing among its citizens.

The increasing significance of incorporating wellbeing indicators into policy decisions is becoming more prominent in New Zealand. Despite this growing importance, there still exists a considerable gap in our understanding of the factors that influence population wellbeing in the country. This knowledge gap is partially attributed to the scarcity of detailed, population-level wellbeing data. The NZ General Social Survey (GSS), a biennial survey of around 9000 individuals<sup>9</sup>, offers wellbeing data across twelve domains: health, housing, income and consumption, jobs and earnings, leisure and free time, knowledge and skills, safety and security, social connections, cultural identity, civic engagement and governance, environmental quality, and subjective wellbeing. Designed based on the NZ Living Standards Framework<sup>10</sup>, which in turn was drawn from the OECD's framework<sup>11</sup>, the GSS lays the foundation for wellbeing assessment in New Zealand. In the context of this study, we focus primarily on the subjective wellbeing domain, focussing on indicators such as life satisfaction, sense of purpose, family wellbeing and mental wellbeing.

Although the GSS sample is considered nationally representative, certain subgroups of the population (that may be of significant policy interest) remain underrepresented due to limitations in sample size. For instance, it is impractical to understand the wellbeing experiences of individuals living in government-sponsored social housing. This is because the number of people who participated in the GSS and are also residents of social housing may be very small. Therefore, to assess the impact of government initiatives targeting this specific population sub-group, comprehensive wellbeing measures applicable to the entire population are needed.

To address this challenge, two strategies offer potential solutions. One approach involves collecting regular wellbeing data for the entire population in a census activity; however, this method is resource-intensive and time-consuming. An alternative approach involves leveraging existing routinely collected data to extrapolate GSS wellbeing measures to the broader population. This may be feasible due to New Zealand's Integrated Data Infrastructure (IDI): a complex database managed by Stats NZ<sup>12</sup>. The IDI contains individual response data (microdata) on people and households, supplemented with anonymised information on education, income, health, justice, and housing. Notably, the IDI facilitates dataset linkage across these areas using a unique identifier variable. Details about this linking process are available elsewhere<sup>13</sup>. Crucially, the IDI houses the GSS data, allowing linkage with the country's Census data which the majority of the nation's population completes (given it is a legal requirement to do so).

The Census is a comprehensive nationwide survey conducted once every five years in New Zealand, with the primary aim of officially counting individuals and households in the country<sup>14</sup>. It also provides a snapshot of various aspects of life, including demographic information, educational qualifications, employment status, and more. Additionally, the Census gathers data on addresses for each household, which are then aggregated at the meshblock level for reporting purposes. A meshblock represents the smallest administrative geographical unit, typically encompassing about 30 to 60 households<sup>15</sup>. Environmental data, such as the extent of green spaces, are also available at the meshblock level and can therefore be linked to the Census data. One notable example is the Healthy Location Index, which captures accessibility to health-promoting elements (e.g., green spaces, physical activity facilities) and health-constraining elements (e.g., alcohol outlets, fast-food shops)<sup>16</sup>. The ability to link such key environmental information to the Census is crucial, given the established links between the environment and wellbeing<sup>17,18</sup>.

The aim of this study is to predict GSS-derived wellbeing measures from Census-based sociodemographic information and meshblock-level environmental indicators. It is important to note that, this study does not make causal claims or explore the determinants of subjective wellbeing; instead, it is purely predictive. If successful, such a predictive model could be used to extrapolate these predicted wellbeing scores to the entire IDI population, thereby creating a population-level estimate of subjective wellbeing. This could yield transformative benefits by facilitating the integration of wellbeing metrics into policy analysis. It also holds the potential to significantly enhance our understanding of how the political, social, and economic landscape impacts the wellbeing and overall functioning of individuals in New Zealand. This would further empower decision-makers to formulate more informed, targeted, and effective policies that address the genuine needs and concerns of New Zealanders.

#### Methods

#### Data sources

The data used in this study was sourced from three datasets: New Zealand General Social Survey (GSS)<sup>9</sup>, New Zealand Census of Population and Dwellings<sup>14</sup>, and the Healthy Location Index (HLI)<sup>19</sup>. Of these, two are present in the New Zealand Integrated Data Infrastructure (IDI), namely the GSS and the Census. All datasets within the IDI are structured as tables in an SQL database and can be linked with one another using the Stats NZ unique identifier variable<sup>20</sup>. All datasets within the IDI can be accessed only from a Stats NZ data laboratory. A formal application to access the IDI datasets, and the IDI data laboratory was submitted and approved by Stats NZ. The methodology used in this research was approved by the AUT University Ethics Committee (AUTEC #21/115).

The study utilized GSS data during the 2018 year, with a sample size of 8,793. More information regarding the GSS and its data collection methodology can be found elsewhere<sup>21,22</sup>. The wellbeing outcome variables, unique identifier variable (snz\_uid) and the meshblock\_code variable were selected from the GSS. The subjective wellbeing outcome variables investigated in this study are listed in Table 1.

Next, the Census 2018 dataset was utilised in this study. Further details about the Census and its methodology are available elsewhere<sup>25</sup>. The size of the dataset was approximately 4.9 million observations with over 300 variables, of which 29 demographic variables were selected as predictors. The choice of these variables was guided by their availability for most of the population. To enhance interpretability, some variables were consolidated

Outcome variable	Survey question in the GSS	Range	Description
1. Life satisfaction	Where zero is completely dissatisfied, and ten is completely satisfied, how do you feel about your life as a whole?	0-10	Completely dissatisfied– Completely satisfied
2. Life worthwhileness	Where zero is not at all worthwhile, and ten is completely worthwhile, overall, to what extent do you feel the things you do in your life are worthwhile?	0-10	Not at all worthwhile– Completely worthwhile
3. Family wellbeing	Where zero means extremely badly and ten means extremely well, how would you rate how your family is doing these days?	0-10	Extremely badly– Extremely well
4. Mental wellbeing	Derived variable, this variable is based on WHO-5's wellbeing index score <sup>23,24</sup>	0-100	Excellent-Poor

 Table 1. GSS wellbeing outcome measures.

into fewer categories due to low counts in some specific categories. Table 2 shows the full list of demographic variables used in the study.

Lastly, data related to the environment was acquired from the Healthy Location Index (HLI) dataset<sup>19</sup>. As this dataset is not present in the IDI, it was imported into the IDI data environment by Stats NZ. The HLI data provides a rank (ranging between 1 and 52,593) for every New Zealand meshblock (excluding oceanic meshblocks). This ranking is determined based on the accessibility of each meshblock (i.e., distance proximity) to both health-promoting features of the environment (e.g., physical activity facilities) and health-constraining features of the environment (e.g., fast-food outlets, takeaway outlets). The methodology involves a straightforward assignment of ranks, offering a transparent depiction of how each meshblock compares in terms of accessibility to these environmental factors. More details about this dataset and the methodology involved in developing this measure can be found elsewhere<sup>16</sup>. A total of 13 environmental variables (shown in Table 3) were used as predictors in this study. All these variables were measured in deciles, ranging from 1 (indicating the highest decile and closest proximity to the environmental feature) to 10 (representing the lowest decile and the farthest distance from the environmental feature).

The GSS dataset was linked with the Census using the unique identifier variable ( $sn_uid$ ) and to the HLI dataset using the meshblock number. After linking these, the dataset underwent a cleaning process to ensure data quality and consistency. Any observations with missing values were removed from the dataset (n=3,135). Unknown or "did not answer" categories in the variables were removed resulting in a data with 5,658 observations and 42 predictor variables (29 Census variables + 13 HLI variables). The demographic distribution of the final dataset (shown in Supplementary table S-1) closely resembles that of the GSS 2018 dataset, indicating a balanced representation of most of the demographic sub-groups without any noticeable over- or under-representation.

#### Modelling

The development of precise predictive models is pivotal in extrapolating GSS data to the broader population. A robust predictive model assists in uncovering patterns within the dataset and establishes a solid foundation for reliable extrapolation. In this study, we employed three distinct predictive models with varying degrees of complexity: (1) Stepwise Linear Regression, (2) Elastic Net Regression, and (3) Random Forest. The modelling process described below was repeated for each of the four wellbeing outcome variables separately (life satisfaction, life worthwhileness, family wellbeing, and mental wellbeing). These models were chosen due to the substantial number of predictor variables (n = 42), and their ability to handle variable selection effectively. Furthermore, the inclusion of Random Forests allowed us to evaluate their ability to model non-linear relationships and complex interactions compared to traditional regression models for predicting subjective wellbeing outcomes.

To begin, the Stepwise Linear Regression method was utilized. It employed an iterative forward and backward selection process to add and remove predictor variables using the Akaike Information Criterion (AIC) as the selection criterion. This ensures that variables were retained or removed based on their joint contribution to model fit, ultimately yielding a subset of relevant variables<sup>27</sup>. While not entirely random, the variable selection process is automated, making it suitable for situations where there are numerous potential predictors. The order of variable selection is determined through statistical criteria rather than random selection. For more detailed information on this model, please refer to Draper and Smith (1998)<sup>28</sup>. Next, we incorporated the Elastic Net Regression model to evaluate its predictive performance in comparison to the Stepwise method. Elastic Net regression provides a unique set of advantages over other regression methods as it is a combination of both Lasso (L1) and Ridge (L2) regularization techniques<sup>29</sup>. This combination facilitates automatic variable selection, enhanced model interpretability and reducing overfitting, making it particularly well-suited for regression tasks involving high-dimensional data<sup>29</sup>. Lastly, we introduced a Random Forest model, to compare its performance against the traditional regression models. The Random Forest is an ensemble learning technique that constructs multiple decision trees and aggregates their predictions to enhance accuracy and reduce overfitting<sup>30</sup>. The Random Forest is effective at handling high-dimensional data as it has inbuilt variable selection, and can capture complex non-linear relationships between variables more effectively than traditional linear regression techniques<sup>31</sup>.

All models were implemented using the *train* function in the 'caret' package in R (version 6.0-94), with the appropriate 'method' argument specified as follows: Stepwise regression: 'glmStepInc', Elastic Net: 'glmnet' (version 4.1-8), and Random Forest: 'rf' (using the randomForest package, version 4.7-1). Furthermore, to mitigate class imbalances inherent within the dataset, class weights were computed as the inverse of the class frequencies and subsequently integrated into the model training process. These weights, operationalised though the 'weights' parameter in the *train* function, serve to recalibrate the model's focus towards underrepresented

Predictor variable	Type of variable	Number of categories	Description
1. Age (in years)	Continuous	NA	0-120
2. Gender	Categorical	2	Male or Female
3. Ethnicity	Categorical	5	European, NZ Māori, Pacific, Asian, Middle Eastern/Latin American/African and Other Ethnic groups
4. Region	Categorical	6	Auckland, Wellington, Northland group (Northland, Bay of Plenty, Gisborne), Rest of North Island, Canterbury, Rest of South Island
5. Marital Status	Categorical	5	Married (not separated), Separated, Divorced or dissolved, Widowed or surviving civil union partner, Never married and never in a civil union
6. Birth Country	Categorical	2	New Zealand, Other
7. Highest Qualification	Categorical	8	No Qualification, School Qualification, Post-school Qualification, Bachelor's degree and Level 7 Qualification, Post-graduate and Honours Degrees, Master's Degree, Doctorate Degree, Overseas Secondary School Qualification
8. Personal Income	Categorical	9	\$0-\$30,000 \$30,001-\$35,000 \$35,001-\$40,000 \$40,001-\$50,000 \$50,001-\$50,000 \$60,001-\$70,000 \$70,001-\$100,000 \$100,001-\$150,000, \$150,001 or More
9. Household Income	Categorical	9	Same as Personal Income
10. Number of income sources	Categorical	5	No source of income, One source, Two sources, Three sources, Four sources, Five or more sources,
11. Workforce Status	Categorical	4	Employed Full-time, Employed Part-time, Unemployed, Not in the Labour Force
12. Study Participation Code	Categorical	3	Full-time study, Part-time study, Not studying
13. Number of Languages spoken	Categorical	7	None, One Language, Two Languages, Three Languages, Four Languages, Five Languages, Six Languages
14. Home Ownership	Categorical	3	Hold in a family trust, Own or partly own, Do not own and do not hold in a family trust
15. Index of Socioeconomic Deprivation Score 2018 <sup>26</sup>	Continuous	Derived variable	823-1552
16. Index of Socioeconomic Deprivation 2018 <sup>26</sup>	Categorical	10	1-Least deprived 10-Most deprived
17. Dwelling dampness indicator	Categorical	4	Always damp, Sometimes damp, Not damp, Don't know
18. Dwelling mould indicator	Categorical	4	Mould over A4 size-always, Mould over A4 size-sometimes, No mould/mould smaller than A4 size, Don't know
19. Difficulty in Seeing	Categorical	4	No difficulty, Some difficulty, A lot of difficulty, Cannot do at all
Continued			

Predictor variable	Type of variable	Number of categories	Description
20. Difficulty in Hearing	Categorical	4	Same as above
21. Difficulty in Washing	Categorical	4	Same as above
22. Difficulty in Communication	Categorical	4	Same as above
23. Difficulty in Remembering	Categorical	4	Same as above
24. Difficulty in Walking	Categorical	4	Same as above
25. Disability indicator	Categorical	2	Not disabled, Disabled
26. Crowding code-based on Canadian National Occupancy Standard	Categorical	5	2 + beds needed, 1 bed needed, no beds needed, 1 bed spare, 2 + beds spare
27. Cigarette smoking behaviour	Categorical	3	Regular Smoker, Ex-Smoker, Never Smoked Regularly
28. Have you ever smoked?	Categorical	2	Yes or no
29. Do you smoke regularly?	Categorical	2	Yes or no

#### Table 2. Predictor variables from the Census 2018 dataset.

Variable	Variable description	Range
1. FruitVeg_rank_dec	Decile of the meshblock rank (ranked based on the distance to the closest Fruit and Veg shop in meters)	1-10
2. Supermarket_rank_dec	Decile of the meshblock rank (ranked based on the distance to the closest supermarket in meters)	1-10
3. PhysicalActivity_rank_dec	Decile of the meshblock rank (ranked based on the distance to the closest physical activity area in meters)	1-10
4. Greenspace_rank_dec	Decile of the meshblock rank (ranked based on the median proximity to greenspace in meters)	1-10
5. Bluespace_rank_dec	Decile of the meshblock rank (ranked based on the median proximity to bluespace area in meters)	1-10
6. Goods_dec	Decile of the sum of meshblock ranks of all environmetal goods (listed above—1 to 5)	1-10
7. FastFood_rank_dec	Decile of the meshblock rank (ranked based on the distance to the closest fast food in meters)	1-10
8. Takeaways_rank_dec	Decile of the meshblock rank (ranked based on the distance to the closest takeaways in meters)	1-10
9. DairyConvenienc_rank_dec	Decile of the meshblock rank (ranked based on the distance to the closest dairy/convenience store in meters)	1-10
10. AlcoholOutlets_rank_dec	Decile of the meshblock rank (ranked based on the distance to the closest alcohol outlet in meters)	1-10
11. GamingVenues_rank_dec	Decile of the meshblock rank (ranked based on the distance to the closest gaming venues in meters)	1-10
12. Bads_dec	Decile of the sum of meshblock ranks of all environmetal bads (listed above—7 to 11)	1-10
13. Env_dec	Decile of the sum of meshblock ranks based on access to access to blue- and greenspace (listed above-4 and 5)	1-10

Table 3. Environment related variables from the Healthy Location dataset (HLI).

.....

classes, thereby improving accuracy in predicting these classes. For instance, a class with substantially fewer instances than others would be assigned a higher weight, incentivising the model to allocate increased computational resources towards accurately predicting instances of this class. This methodological adjustment is crucial in fostering a balanced predictive performance, counteracting the model's inherent propensity to bias predictions in favour of overrepresented classes.

Firstly, the dataset was split into a training set and a testing set in a 70:30 ratio. The training set, consisting of 70% of the data (n = 3963 observations), was further subjected to a tenfold cross-validation process to evaluate and select the best model parameters. During this cross-validation process, various combinations of hyperparameters (e.g., mtry and ntree values for the random forest model) were evaluated, and the optimal values (that yielded the lowest root mean squared error) were used to train the final model. The final models were trained on the entire training dataset using these optimal parameters. The performance of the final models was then evaluated on the testing dataset (n = 1695 observations) to assess their predictive capabilities and generalisation to unseen data. For the Random Forest, the importance of each variable for improving model performance was estimated using the varImp function in the 'caret' R package, which evaluates the contribution of each predictor to the overall predictive performance of the model. Specifically, variable importance is assessed based on the mean decrease in accuracy or Gini impurity when a variable is excluded or permuted. Two variations of the model were fit, one incorporating environment-related variables from the HLI dataset, and another excluding HLI indicators. This was performed to examine how environmental data affected the predictive performance of each model. The performance of all models were assessed using root mean squared error (RMSE), mean absolute error (MAE), and R-squared  $(R^2)$ . As a further check, the Pearson's correlation between the observed and predicted values were also evaluated.

		Predicted mean ± SD		
Outcome variable	Observed mean ± SD	Stepwise regression	Elastic net regression	Random forest
Life satisfaction	$7.75 \pm 1.65$	$6.23 \pm 1.48$	$6.09 \pm 1.19$	$7.76 \pm 0.23$
Life worthwhileness	$8.15 \pm 1.54$	$6.68 \pm 1.42$	6.61±1.35	8.14±0.21
Family wellbeing	$7.86 \pm 1.61$	$6.69 \pm 1.97$	$6.67 \pm 1.65$	$7.87 \pm 0.21$
Mental wellbeing	63.16±17.75	$55.41 \pm 13.58$	$54.55 \pm 10.57$	63.23±2.32

**Table 4**. Descriptive statistics (obtained from the testing dataset, n = 1695) for observed and predictedwellbeing variables.

	Outcome variable	RMSE	MAE	<b>R</b> <sup>2</sup>
Stepwise regression	Life satisfaction	2.534	2.081	0.028
	Life worthwhileness	2.463	2.012	0.013
Stepwise regression	Family wellbeing	2.683	2.154	0.010
	Mental wellbeing	21.509	17.459	0.040
	Life satisfaction	2.497	2.085	0.028
Elastic net regression	Life worthwhileness	2.461	2.030	0.015
	Family wellbeing	2.478	2.008	0.013
	Mental wellbeing	20.325	16.772	0.055
Random forest (with environmental variables)	Life satisfaction	1.595	1.200	0.077
	Life worthwhileness	1.508	1.164	0.045
	Family wellbeing	1.583	1.195	0.040
	Mental wellbeing	17.226	13.715	0.062
	Life satisfaction	1.596	1.198	0.072
Random forest	Life worthwhileness	1.505	1.165	0.050
(without environmental variables)	Family wellbeing	1.582	1.198	0.040
	Mental wellbeing	17.273	13.745	0.053

Table 5. Model performance metrics.

#### Code availability

Code associated with this study is available as a supplementary file. However, given that the analysis was carried out within the Data lab environment, the models are not publicly available. For more details, please refer to the Data Availability section.

#### Results

We employed three distinct models to predict four wellbeing variables: life satisfaction, life worthwhileness, family wellbeing, and mental wellbeing. Table 4 provides a summary of both the observed mean and standard deviation, alongside the predicted values for all models. Notably, the Random Forest model exhibited superior performance, with predictions that were closely aligned with the observed values. These results were obtained through the evaluation of model performance on the test dataset, comprising 30% of the original dataset (n = 1695).

Table 5 provides an overview of the performance metrics for all predictive models. Notably, the Random Forest models demonstrated stronger performance with lower RMSE (ranging between 1.5 and 1.6 for life satisfaction, life worthwhileness, and family wellbeing). However, the R-squared ( $R^2$ ) values were relatively low (~0.006), suggesting that these models had limited explanatory capabilities. Traditional models (Stepwise regression and Elastic Net) produced higher RMSE values (~2.5) and even lower  $R^2$  values (<0.003) for these wellbeing variables. Table 5 displays the results with and without the inclusion of environmental variables for the Random Forest model only (given this was the best performing). The incorporation of environmental features had a negligible impact on the model's predictive capacity. Furthermore, we assessed the correlation between the observed and predicted values produced by the Random Forest model (without environmental variables). This correlation ranged from weak to moderate, falling within the range of 0.202 to 0.250, for all wellbeing outcome variables (all p < 0.05). Supplementary table S-2 shows the importance of the top 10 predictor variables employed by the random forest model.

#### Discussion

The primary aim of this study was to evaluate the predictive efficacy of population-level socio-demographic variables in predicting GSS-based subjective wellbeing outcomes, encompassing life satisfaction, life worthwhileness, family wellbeing, and mental wellbeing. This analysis was augmented by incorporating environmental data from the Healthy Location Index. The study employed three distinct predictive models:

Stepwise Regression, Elastic Net, and Random Forest. Our results demonstrated the models' ability to predict wellbeing outcomes, as evidenced by their low RMSE values, by utilizing a concise set of easily accessible socio-demographic variables from the Census. However, the low R<sup>2</sup> values suggest a constrained capacity to account for the extensive variability in the dependant variables. In practical terms, while the models are adept at approximating group-level averages with reasonable precision —an approach relevant for policy applications— they fail to capture the underlying dynamics or variance in individual-level wellbeing outcomes, which is critical for tailoring interventions and understanding subjective wellbeing in depth. This limitation may be influenced by various factors, including dataset characteristics, as discussed in subsequent sections. Notably, this aligns with findings from Lundberg et al. [2024] and Salganik et al. [2020], which highlight that even advanced models may struggle to explain the variability in subjective outcomes due to irreducible error<sup>32,33</sup>. While our findings emphasise the need for further improvements in predictive modelling, they also underscore the fundamental limits of explainability for subjective and multidimensional outcomes due to the complex and dynamic nature of human lives.

In our investigation, Random Forest models outperformed conventional modelling techniques like Elastic Net and Stepwise regression in terms of predictive capability. This may be because random forest algorithms are capable of capturing complex nonlinear relationships in the data, handling multicollinearity, and reducing overfitting through their ensemble nature<sup>31,34</sup>. While previous studies have employed similar methodologies to predict clinical outcomes such as the incidence of cardiovascular diseases and other chronic conditions<sup>35</sup>, our study stands out by predicting subjective wellbeing outcomes (e.g., life satisfaction) by utilising a straightforward demographic variable set.

The inclusion of environmental variables from the HLI dataset did not result in a significant improvement in model performance when compared to models that solely relied on socio-demographic factors. Yet, these environmental variables ranked among the top 10 important predictors when assessed using the *varImp* function. This suggests that while environmental factors are associated with subjective wellbeing outcomes, they do not necessarily have any causal influence. Prior research has indicated a connection between the HLI indicators and deprivation<sup>16</sup>, primarily determined using various socio-demographic indicators such as education, income, and housing data from the Census. Given that we have already included a range of these Census-level sociodemographic variables in our analyses, the inclusion of environmental variables may not have offered any additional insights beyond what we had already captured through the Census data.

Additionally, it's worth noting that the environmental variables from the HLI dataset primarily capture proximity to various environmental elements but do not consider the total number, variety, or quality of such facilities. It is known that overall extent and quality of green/blue space within an area is related to mental health<sup>36,37</sup>, and studies have established the importance of environmental factors in influencing an individual's mental health<sup>38,39</sup>. It should also be noted that HLI is an area-level measure, yet we were predicting individual-level outcomes, which could have also attenuated the effect of the environment. Future studies could explore the utility of a more nuanced selection of environmental variables in the modelling process.

Although our predictions were reasonable, there are limitations in our approach that should be discussed. Firstly, the wellbeing data from the GSS 2018 dataset used to train the models did not have a uniform distribution of responses across the measurement scale. For instance, the outcome variable 'life satisfaction' ranged from 1 to 11, and over 50% of respondents reported a score of either 7 or 8. This imbalance may be inherent to the subjective nature of the question. Despite incorporating weights into the model training process, the majority of our predictions tended to cluster around scores of 7 and 8. Since this range of values closely aligns with that of the observed values, the models achieved a relatively low RMSE (<1.6). However, a lower correlation between the observed and predicted values (0.20-0.25) suggests that the model predictions within this narrow range were not linearly associated with the observed data. This discrepancy can likely be attributed to the limited range of values present in the GSS dataset. It's also worth noting that while our predictions typically fell within a 1-2-point range of the true scores, this apparent accuracy could be misleading. This is because the true scores themselves predominantly fell within this same 1-2-point range, and consequently, the proportional error is relatively high. While this study did not include explicit uncertainty quantification, future work could employ methods such as bootstrap resampling to estimate confidence intervals for predictions. These techniques could provide additional insights into the variability and robustness of model outputs, particularly in contexts where subjective outcomes are clustered within a narrow range.

Understanding subjective wellbeing, especially when collected through surveys, is complex. Unlike quantifying tangible health conditions (e.g., cardiovascular disease, obesity, diabetes), subjective wellbeing relies on self-reported responses, which can vary based on how an individual interprets the question. For example, two people who choose scores of 7 might perceive those scores differently. Moreover, a lower score might not necessarily indicate less satisfaction relative to another person, it could reflect an individual's unique understanding of the scale. Without a benchmark for validation, it is challenging to confidently interpret model results. Another important consideration is that these outcome scores reflect an individual's overall wellbeing experience over time, not just their feelings on the day of the survey. However, someone generally satisfied with life might choose a lower score if recent unpleasant events influenced their mood. The subjective nature of these outcomes makes their validation difficult.

Another limitation arises from the dataset cleaning process, particularly the exclusion of nearly 3% of the Māori population due to missing values (see Supplementary table S-1). This exclusion could have potentially introduced bias into the model's predictions and overall outcomes. Similarly, another limitation pertains to the Census 2018 dataset which had a lower response rate than expected. To address this challenge, Stats NZ employed alternative strategies to impute missing data. These strategies involved leveraging other available microdata within the Integrated Data Infrastructure (IDI) to fill in the gaps and enhance the completeness of the dataset. Although the data imputation process is beneficial, it introduces a potential source of bias or uncertainty

in our results, as the imputed values may not accurately capture the true characteristics of the non-respondent population. Further information regarding this issue can be explored in "2018 Census collection response rates unacceptably low" by Stats NZ (2018)<sup>40</sup>.

To enhance the predictive performance for future studies, we recommend exploring additional analyses, improved data handling, and engaging in alternate feature engineering strategies. For instance, while the re-weighting strategy used in this study aims to address the underrepresentation of certain classes in the GSS dataset, it does not eliminate biases inherent in the original data, potentially leading to the replication of these biases in model predictions. An alternative approach, such as bootstrapped re-sampling<sup>41</sup>, could explore the impact of synthetic samples on prediction. Future work should compare re-sampling techniques, synthetic data generation, and augmentation methods to address class imbalance and underlying biases more effectively while preserving model robustness. Additionally, this study addressed missing data by removing incomplete observations to maintain dataset consistency. While this approach is straightforward, it may have introduced bias by excluding certain groups of respondents. Future work could explore alternative methods, such as imputation techniques, to better understand the potential impact of missing data on model performance and ensure the robustness of findings.

Next, a broader range of demographic variables could be considered to provide a more comprehensive representation of individual characteristics. One area for future exploration could be examining how different treatments of the outcome variable impact the model's predictive accuracy. For instance, our model used life satisfaction as a continuous variable, rather than categorizing it (e.g., low, medium, high). However, establishing thresholds for these classes could be uncertain and may require guidance from industry experts. Additionally, creating composite indices that capture multiple dimensions of wellbeing, or integrating other data sources available in the IDI (e.g., health data) as predictors could potentially lead to improved model performance.

While the IDI and Census provide rich, granular datasets, their reliance on periodic collection—such as the five-year interval for census data in New Zealand—presents a fundamental limitation for generating real-time or frequently updated predictions. Addressing this limitation may involve exploring complementary data sources, such as observation (EO) data, which have been effectively combined with machine learning to estimate health and living conditions<sup>42–44</sup>. Lastly, considering alternative modelling techniques beyond the ones explored in this study, such as neural networks or support vector machines, may provide further insights into predicting wellbeing outcomes and improve model performance.

#### Conclusion

Our findings indicate that a Random Forest model, in conjunction with census-level socio-demographic variables, yields moderate predictive efficacy for a range of GSS-based subjective wellbeing measures. This outcome underscores the potential of this methodological approach. However, it is imperative to acknowledge limitations arising from the subjective nature and distribution characteristics of the outcome variables. While our study offers valuable insights into predicting wellbeing outcomes using predictive modelling techniques, there is significant scope for improvement. By refining the modelling approach, incorporating more diverse data sources (e.g., health records within the IDI), and employing advanced analytical methods (e.g., deep learning), future research can contribute to a more accurate and comprehensive understanding of population wellbeing and offer robust tools for evidence-based policymaking.

#### Data availability

The data that support the findings of this study are available from Stats NZ'S Integrated Data Infrastructure, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. However, access to the data can be obtained upon approval by Stats NZ. To initiate the process, a formal application must be submitted to the Stats NZ Microdata team. For step-by-step guidance on how to request access, please refer to the following link: https://www.stats.govt.nz/integrated-data/apply-to-us e-microdata-for-research. Please get in touch with the corresponding author for more details.

Received: 15 April 2024; Accepted: 17 February 2025 Published online: 25 February 2025

#### References

- Huppert, F. A. & So, T. T. C. Flourishing across Europe: Application of a new conceptual framework for defining well-being. Soc. Indic. Res. 110(3), 837–861 (2013).
- 2. Coscieme, L. et al. Overcoming the myths of mainstream economics to enable a new wellbeing economy. *Sustainability* 11(16), 4374 (2019).
- 3. Treasury, N. Z. (ed.) The Wellbeing Budget 2019 (Wellington, 2019).
- 4. Jahoda, M., *Current concepts of positive mental health*. Current concepts of positive mental health. Basic Books. xxi, 136-xxi, 136 (1958).
- 5. Seligman, M.E.P., Authentic Happiness: Using the New Positive Psychology to Realize Your Potential for Lasting Fulfillment (Free Press. xiv, 321-xiv, 321, 2002).
- Diener, E., Suh, E. M., Lucas, R. E. & Smith, H. L. Subjective well-being: Three decades of progress. *Psychol. Bull.* 125(2), 276–302 (1999).
- 7. Seligman, M. E. P. Flourish: A Visionary New Understanding of Happiness and Well-Being (Simon and Schuster, 2012).
- 8. Thompson, S. & Marks N. Measuring Well-Being in Policy: Issues and Applications. New Economics Foundation (2008).
- 9. Stats, N. Z., New Zealand General Social Survey 2018, Statistics (eds NZ). (Wellington, 2018).
- 10. Treasury, N. Z. The Treasury Approach to the Living Standards Framework (eds. Treasury N) (Wellington, 2018).
- 11. OECD, OECD Guidelines on Measuring Subjective Well-Being (2013).
- 12. Stats NZ. Integrated Data Tools. Available from https://www.digital.govt.nz/showcase/integrated-data-tools/.
- 13. Kvalsvig, A., Gibb, S. & Teng, A. Linkage error and linkage bias: A guide for IDI users. University of Otago (2019).

- 14. Stats NZ. Census. Available from https://www.stats.govt.nz/topics/census.
- 15. Stats, N. Z. Statistical standard for meshblock (Wellington, 2016).
- 16. Marek, L. et al. The good, the bad, and the environment: developing an area-based measure of access to health-promoting and health-constraining environments in New Zealand. Int. J. Health Geogr. 20, 1-20 (2021).
- 17. Gong, Y. et al. A systematic review of the relationship between objective measurements of the urban environment and psychological distress. Environ. Int. 96, 48-57 (2016).
- 18. Engemann, K. et al. Residential green space in childhood is associated with lower risk of psychiatric disorders from adolescence into adulthood. Proc. Natl. Acad. Sci. 116(11), 5188-5193 (2019).
- 19. Hobbs, M. et al. Unhealthy environments are associated with adverse mental health and psychological distress: Cross-sectional evidence from nationally representative data in New Zealand. Prev. Med. 145, 106416 (2021).
- 20. Milne, B. J. et al. Data resource profile: the New Zealand integrated data infrastructure (IDI). Int. J. Epidemiol. 48(3), 677e (2019). 21. Stats NZ. General Social Survey (GSS). Available from https://datainfoplus.stats.govt.nz/Item/nz.govt.stats/2ed50ad6-8ab8-47df-8 83d-210a51b50043#:~:text=The%20GSS%20uses%20a%20three,characteristics%20of%20the%20whole%20country.
- 22. Stats NZ. About the General Social Survey. 2018; Available from https://www.stats.govt.nz/help-with-surveys/list-of-stats-nz-surve s/about-the-general-social-survey/.
- 23. Staehr, J. K. The use of well-being measures in primary health care-the DepCare project. In World Health Organization, Regional Office for Europe: Well-Being Measures in Primary Health Care-the DepCare Project. Geneva: World Health Organization (1998). 24. Topp, C. W., Østergaard, S. D., Søndergaard, S. & Bech, P. The WHO-5 Well-Being Index: A systematic review of the literature.
- Psychother. Psychosom. 84(3), 167-176 (2015).
- 25. Stats NZ. Census. 2023; Available from https://www.census.govt.nz/.
- 26. Atkinson, J., Salmond, C. & Crampton, P. NZDep2018 Index of Deprivation, Interim Research Report, 65 (Wellington, 2019).
- 27. Miller, A. Subset Selection in Regression (CRC Press, London, 2002).
- 28. Draper, N. R. & Smith, H. Applied Regression Analysis Vol. 326 (Wiley, 1998).
- 29. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B: Stat. Methodol. 67(2), 301-320 (2005)
- 30. Breiman, L. Random forests. Mach. Learn. 45, 5-32 (2001).
- 31. Liaw, A. & Wiener, M. Classification and regression by randomForest. R news 2(3), 18-22 (2002).
- 32. Lundberg, I. et al. The origins of unpredictability in life outcome prediction tasks. Proc. Natl. Acad. Sci. 121(24), e2322973121 (2024)
- 33. Salganik, M. J. et al. Measuring the predictability of life outcomes with a scientific mass collaboration. Proc. Natl. Acad. Sci. 117(15), 8398-8403 (2020).
- 34. Cutler, D. R. et al. Random forests for classification in ecology. Ecology 88(11), 2783-2792 (2007).
- 35. Luo, W. et al. Is demography destiny? Application of machine learning techniques to accurately predict population health outcomes from a minimal demographic dataset. PLoS One 10(5), e0125602 (2015).
- 36. Nutsford, D., Pearson, A. L. & Kingham, S. An ecological study investigating the association between access to urban green space and mental health. Public Health 127(11), 1005-1011 (2013).
- 37. Nutsford, D., Pearson, A. L., Kingham, S. & Reitsma, F. Residential exposure to visible blue space (but not green space) associated with lower psychological distress in a capital city. Health & Place 39, 70-78 (2016).
- 38. Evans, G. W. The built environment and mental health. J. Urban Health 80(4), 536-555 (2003).
- 39. Hartig, T., Mitchell, R., De Vries, S. & Frumkin, H. Nature and health. Ann. Rev. Public Health 35, 207-228 (2014).
- 40. Stats NZ. 2018 Census collection response rates unacceptably low. 2018; Available from: https://www.stats.govt.nz/methods/2018-c ensus-collection-response-rates-unacceptably-low.
- 41. Berrar, D. & Dubitzky, W. Bootstrapping. In Encyclopedia of Systems Biology (eds Dubitzky, W. et al.) 158-162 (Springer New York, 2013).
- 42. Daoud, A. et al. Using satellite images and deep learning to measure health and living standards in India. Soc. Indic. Res. 167(1), 475-505 (2023)
- 43. Pettersson, M. B., et al. Time series of satellite imagery improve deep learning estimates of neighborhood-level poverty in Africa. In International Joint Conference on Artificial Intelligence. Macao, China (2023).
- 44. Sakamoto, K., Jerzak, C. T. & Daoud, A. A scoping review of earth observation and machine learning for causal inference: Implications for the geography of poverty. arXiv preprint arXiv:2406.02584 (2024).

#### Acknowledgements

AN received the AUT Vice Chancellor's Doctoral Scholarship Award which funded the study. The results of this study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation.

#### Disclaimer

These results are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI) which is carefully managed by Stats NZ. For more information about the IDI please visit https://www.stats.govt.nz/integrated-data/.

#### Author contributions

AN and TS was responsible for study design with assistance from SD and GP. Data cleaning and processing were performed by AN and assistance from TS. AN performed the analysis and drafted the manuscript, with critical feedback provided by all authors.

#### Declarations

#### Competing interests

The authors declare no competing interests.

#### Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/1 0.1038/s41598-025-90852-0.

Correspondence and requests for materials should be addressed to T.S.

Reprints and permissions information is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025