Multiple Pathways to Regional Well-Being

A data-driven analysis of how different well-being profiles influence life satisfaction, and how partial orderings uncover complexity beyond standard rankings

Micol Colombo, mat. 874860
Daria Marinucci, mat. 879727
Leonardo Roman, mat. 886288
Matteo Suardi, mat. 930935

Academic Year 2024/2025

Contents

Chapter 1

Introduction, Research Question and Methodology

1.1 Introduction

Over the past decades, international institutions and scholars have increasingly moved beyond GDP-centric perspectives in assessing societal progress. The recognition that well-being is a multidimensional construct—encompassing both material conditions and relational aspects of life—has led to the development of more comprehensive frameworks. Among these, the OECD Regional Well-Being initiative provides one of the most structured and granular tools for capturing the multifaceted nature of quality of life across 447 subnational regions in 38 member countries.

The OECD dataset includes 11 dimensions of well-being: Income, Jobs, Housing, Education, Health, Safety, Civic Engagement, Environment, Accessibility to Services, Community, and Life Satisfaction. These indicators are reported on a standardized 0–10 scale, calculated using percentile-based normalization methods with trimming of extreme values (below the 4th and above the 96th percentiles). The resulting scores ensure comparability across regions and minimize distortion caused by outliers.

Crucially, the dataset includes both objective indicators—such as mortality rates, unemployment, broadband access, or air pollution—and subjective measures, most notably self-assessed life satisfaction and perceived social support. This dual nature allows for the investigation of both structural and experiential determinants of well-being, as well as the interdependencies between them.

However, the potential of this data is often undermined by the widespread use of composite indices and rankings, which condense multidimensional well-being into a single metric. Such practices tend to obscure the diversity of regional profiles and the possibility that different configurations of strengths and weaknesses may result in similar subjective outcomes. For instance, a region characterized by high income and poor health might achieve the same life satisfaction score as one marked by moderate income and strong social capital.

Against this backdrop, the present project adopts a data-driven, multilevel approach to investigate the determinants of life satisfaction across OECD regions. It aims to uncover whether multiple, qualitatively distinct pathways to well-being exist and how they vary across territorial contexts.

To address this aim, the project integrates several complementary methodologies:

- **Exploratory Data Analysis (EDA)** to examine distributional properties, miss-

ingness patterns, and correlation structures;

- **Supervised machine learning models** (Random Forest and XGBoost) to predict life satisfaction based on both standardized scores and raw indicators;
- **SHAP (SHapley Additive exPlanations)** to interpret model predictions and quantify the contribution of each feature;
- **KMeans clustering and PCA visualization** to detect alternative well-being configurations and evaluate their subjective outcomes;
- **Policy simulation** to assess the hypothetical impact of targeted interventions on predicted life satisfaction.

This methodological framework is applied separately to two distinct but complementary datasets: one based on the normalized OECD well-being scores, and the other on the original structural indicators. This dual-track strategy allows for both dimension-level interpretability and indicator-level granularity, enhancing the robustness and scope of the analysis.

1.2 Research question and analytical strategy

The project is structured around the following overarching research question:

> **What determines life satisfaction in OECD regions, and how can we explain and predict its variation across different territorial contexts?**

Life satisfaction serves as the target variable throughout the study, representing a comprehensive, subjective assessment of well-being. Unlike previous literature focused primarily on individual-level data, this research centers on regional-level determinants—such as infrastructure, environment, public safety, education systems, and social cohesion—that shape the conditions in which people live.

To provide a comprehensive answer to the research question, the analysis is organized into two main blocks:

1.2.1 Predictive and interpretative modeling

This component seeks to quantify how much life satisfaction can be predicted from structural well-being indicators, and which features contribute most significantly to these predictions. Key objectives include:

- Comparing the predictive power of Random Forest and XGBoost regressors;
- Applying SHAP values to interpret global and local feature importance;
- Investigating whether the same determinants hold across the score-based and indicator-based models;
- Simulating policy scenarios to evaluate potential improvements in life satisfaction.

1.2.2 Structural diversity and pathways to well-being

This second block explores whether regions can reach comparable life satisfaction levels through qualitatively different configurations of structural indicators. It addresses:

- Whether clusters of regions emerge with distinct profiles but similar subjective outcomes;
- How social and economic trade-offs (e.g., strong community vs. high income) influence well-being;
- Whether equifinality—a condition where different paths lead to the same outcome—can be empirically demonstrated;
- The policy implications of recognizing and addressing this structural diversity.

Through this combined approach, the project not only predicts life satisfaction with high accuracy but also provides a theoretically informed, empirically grounded understanding of the plural nature of regional well-being.

1.3 Methodological overview

All analyses are conducted using Python (with libraries including `pandas`, `scikit-learn`, `xgboost`, and `shap`) and R (for exploratory data analysis), ensuring a high level of computational transparency and reproducibility. Data cleaning and preprocessing include imputation K-Nearest Neighbors imputation. Modeling results are rigorously evaluated using cross-validation, performance metrics (RMSE, MAE, R^2), and statistical significance tests.

Clustering is implemented via KMeans on z-standardized dimensions, and results are interpreted through cluster profiling, life satisfaction overlays, and PCA-based spatial visualization.

Overall, the methodological pipeline integrates machine learning, explainable AI, and unsupervised learning into a cohesive framework capable of addressing both explanatory and prescriptive dimensions of regional life satisfaction.

Chapter 2

Exploratory Data Analysis (EDA)

2.1 EDA on the Score dataset

The `Score_Last` dataset contains aggregated scores for eleven dimensions of well-being, each normalized on a 0–10 scale, for 447 subnational regions across 38 OECD countries. These variables include `Education`, `Jobs`, `Income`, `Safety`, `Health`, `Environment`, `Civic engagement`, `Accessibility to services`, `Housing`, `Community`, and the dependent variable `Life satisfaction`.

2.1.1 Data cleaning and preprocessing

Initially, all variables were encoded as strings and required coercion into numerical types. During this process, non-numeric values were converted to `NA`. Diagnostics performed with `skim_without_charts()` revealed that most variables had completeness levels exceeding 0.9, with the exception of `Income`, `Housing`, and `Community`, which exhibited higher rates of missingness.

2.1.2 Two-step imputation strategy

To address these missing values, a hierarchical imputation strategy was adopted. First, the mean of each variable was calculated within each country group to preserve national characteristics. Where country-level values were entirely missing for a variable, the global median was used as a fallback. This strategy ensured full data coverage and maintained interpretability and comparability across regions.

2.1.3 Distributional analysis

Histograms and boxplots revealed that most well-being dimensions were unimodally distributed with slight left skewness, indicative of generally favorable conditions.`Safety` was a notable outlier, displaying an extreme peak near the upper boundary, while `Income` showed a long right tail. `Civic engagement` exhibited a bimodal distribution, suggesting distinct regional patterns in participation.

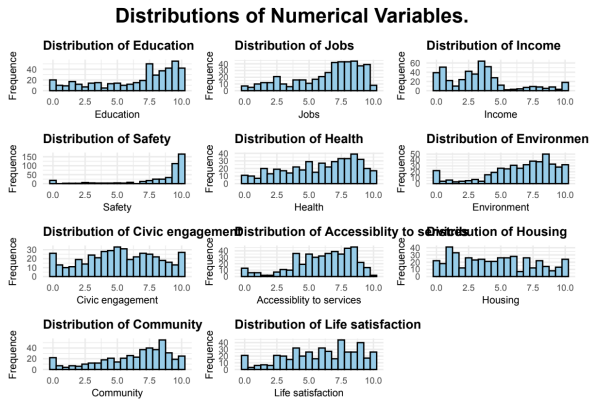


Figure 2.1: Distributions of numerical variables in the Score dataset

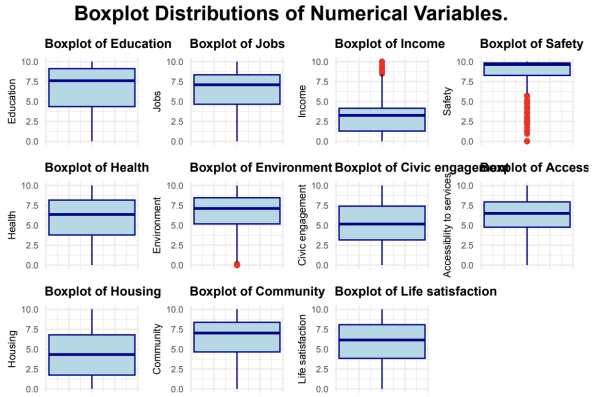


Figure 2.2: Boxplot distributions of numerical variables in the Score dataset

2.1.4 Correlation analysis

The Pearson correlation matrix showed strong positive associations among most variables. `Life satisfaction` correlated most strongly with `Community` ($r = 0.63$), `Housing`, `Income`, `Jobs`, and `Environment`. Scatterplots illustrated these relationships as generally monotonic, with indications of diminishing returns at higher values.

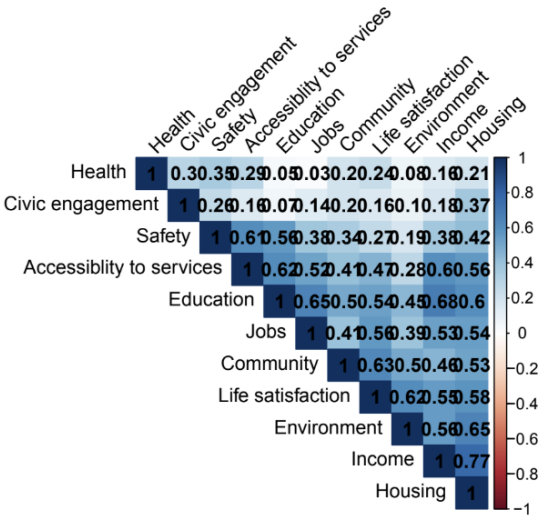


Figure 2.3: Correlation matrix of the numerical variables

2.1.5 Geospatial aggregation

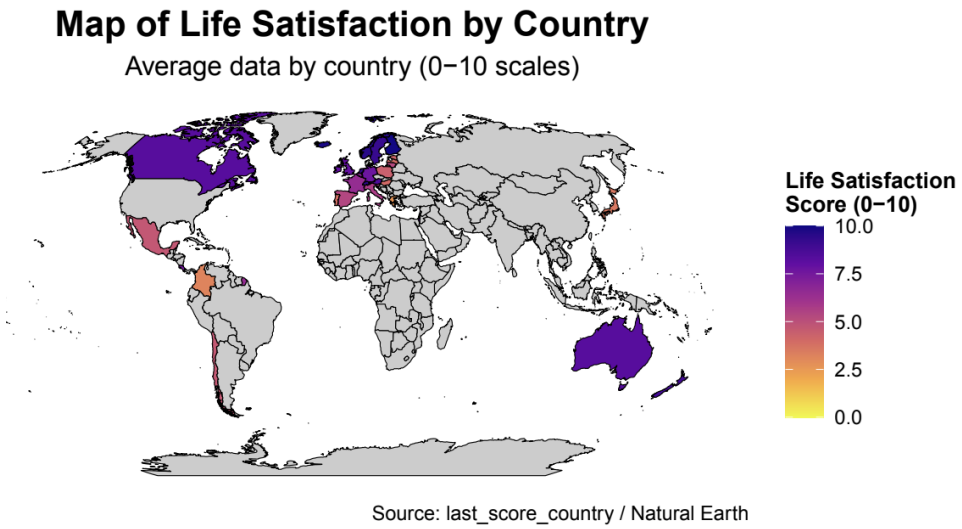


Figure 2.4: Life satisfaction for each Country

Aggregating regional scores at the national level and visualizing them through choropleth and bubble maps revealed clear geographic gradients. High satisfaction scores clustered in Nordic countries and Oceania, while lower values were observed in Turkey, Colombia, and Greece. These patterns aligned with known differences in welfare models and socio-political structures.

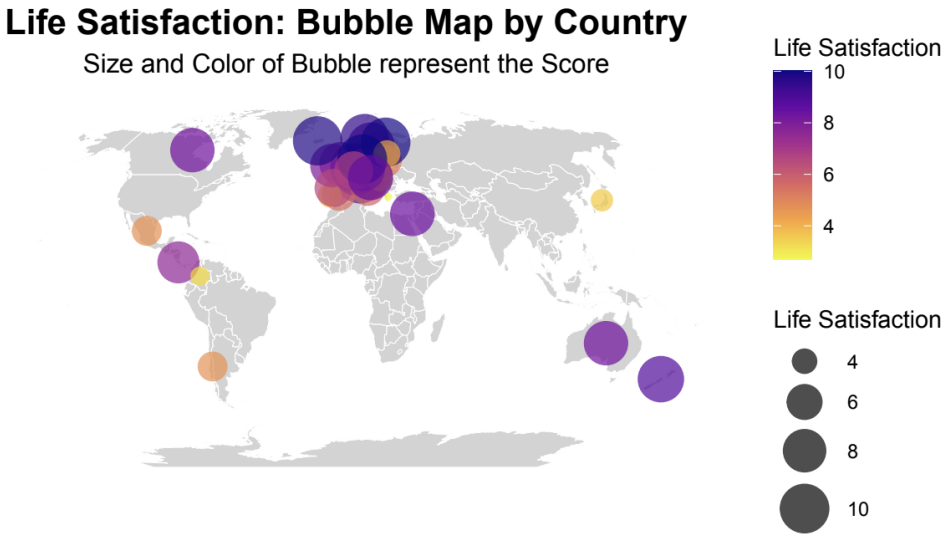


Figure 2.5: Bubble map for Life satisfaction

2.2 EDA on the Indicator dataset

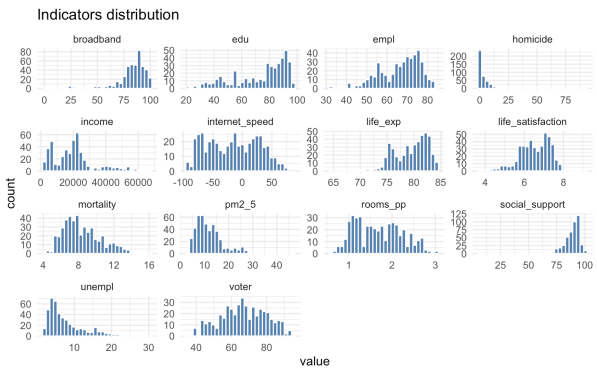
The `Indicator_Last_Region.csv` dataset contains 14 numeric indicators representing objective measures of well-being, such as `education`, `income`, `mortality`, `pm2_5` (air pollution), and `life_satisfaction`, for the same set of 447 regions.

2.2.1 Data cleaning and variable transformation

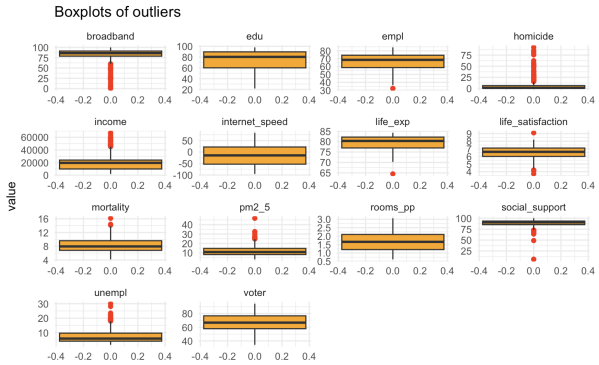
Variables were renamed to ensure clarity and consistency. Missing values were addressed using the same two-step imputation strategy described above: within-country means followed by global medians. This allowed the retention of all regions in subsequent analysis.

2.2.2 Distributional characteristics

Descriptive statistics and histograms revealed strong heterogeneity. Income, mortality, and pollution indicators exhibited long right tails, while broadband access and social support were more symmetrically distributed. `Life satisfaction` followed a near-normal distribution with slight left skewness.



Boxplots confirmed the presence of extreme outliers in several variables.



2.2.3 Correlation and partial correlation analysis

Pearson correlations highlighted strong associations among education, income, and employment. `Life satisfaction` was positively associated with community, income, and rooms per person, and negatively associated with mortality and PM2.5. Partial correlation analysis reduced the strength of some of these associations, revealing the mediating role of latent factors. For example, the employment-income link weakened after adjustment, while the mortality-life expectancy relationship remained strong.

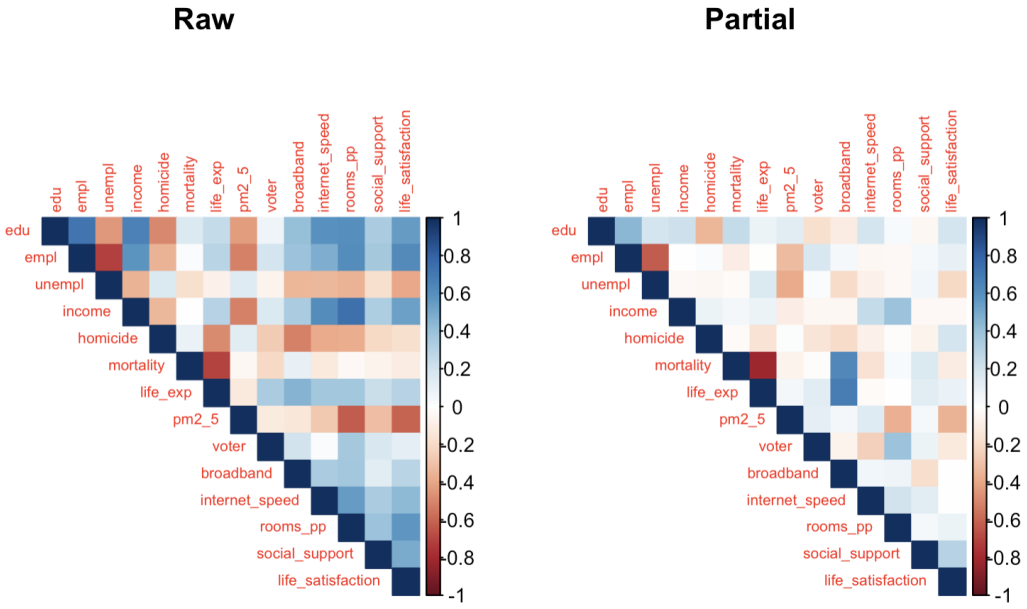


Figure 2.6: Comparison between raw and partial correlations

2.2.4 Multicollinearity and latent structures

Scatterplot matrices confirmed high multicollinearity among socioeconomic indicators. A latent "socioeconomic capital" structure emerged, encompassing education, income, employment, and housing. Negative associations were observed with unemployment and pollution. These findings indicated the need for dimensionality reduction and penalized modeling approaches.

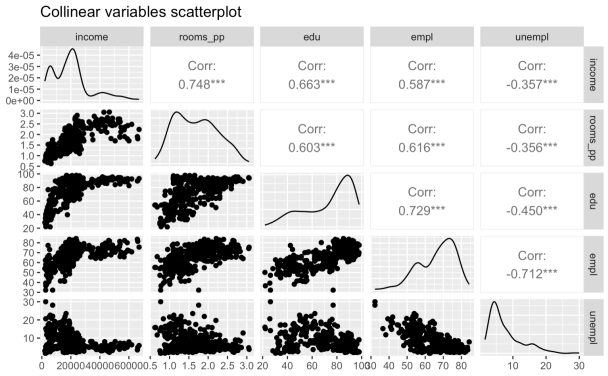


Figure 2.7: Collinear variables scatterplot

2.2.5 Cross-country aggregation and comparative rankings

Regional indicators were averaged at the country level, inverted where necessary, and standardized using z-scores. Visualizations of the top performers confirmed that countries such as Finland and Switzerland consistently led across multiple domains, while Colombia and Turkey ranked lowest. Interesting deviations were observed for countries like Costa Rica, which scored high on life satisfaction despite modest economic conditions.

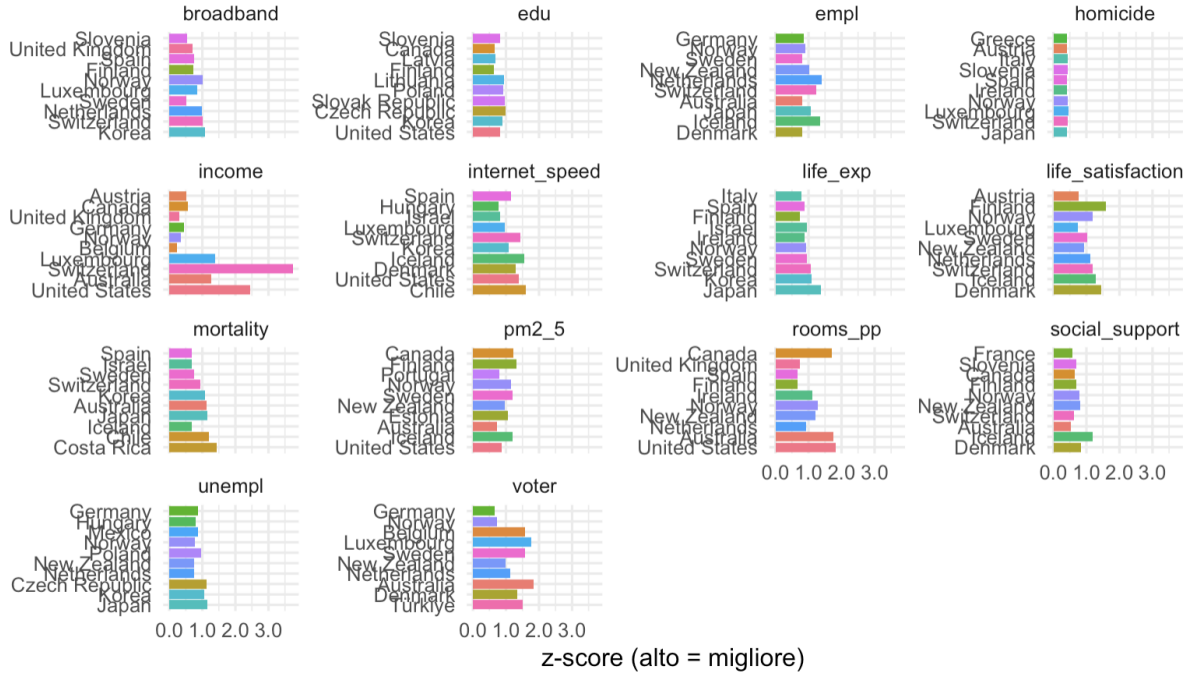


Figure 2.8: Top 10 Countries for each indicator (z-score, inverted negatives)

2.2.6 Bivariate insights on life satisfaction

Scatterplots revealed that income, social support, and education had positive but dispersed relationships with life satisfaction. Mortality and air pollution showed mild to moderate negative associations. These patterns supported the idea of life satisfaction as an emergent property of multiple interrelated factors.

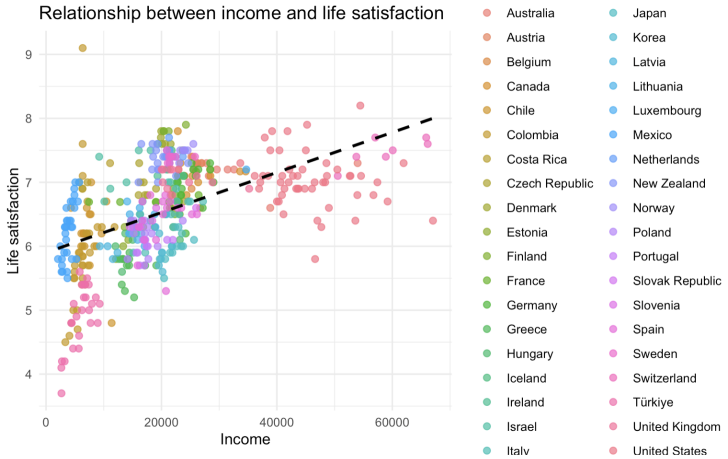


Figure 2.9: Scatterplot for income and life satisfaction regression

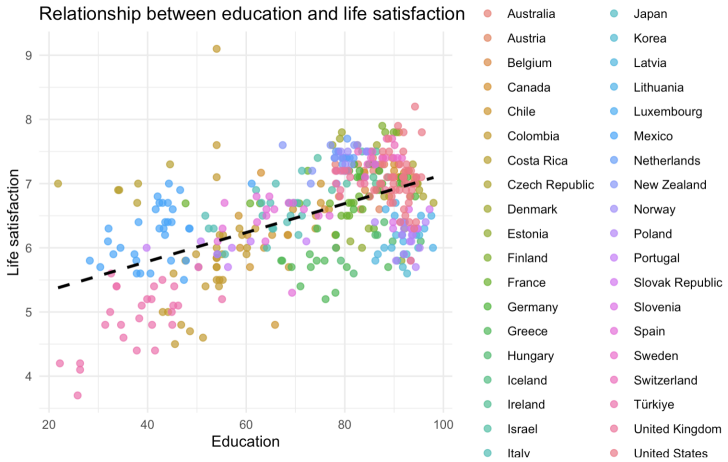


Figure 2.10: Scatterplot for education and life satisfaction regression

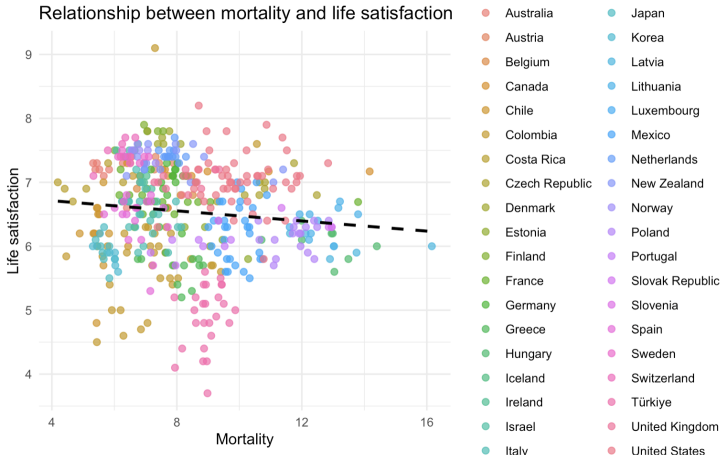


Figure 2.11: Scatterplot for mortality and life satisfaction regression

2.3 Comparative reflection and transition to KNN imputation

While the exploratory analyses for both datasets revealed consistent patterns—such as the salience of income, housing, and community—they also showed dataset-specific dynamics. The Score dataset emphasized normalized, high-level measures, whereas the Indicator dataset provided granular, domain-specific metrics. Despite methodological similarities, such as the two-step imputation approach and correlation-based diagnostics, the complexity of real-world data called for improved techniques.

To enhance model performance and robustness, we transitioned from the hierarchical imputation strategy to a non-parametric, distance-based approach.

2.4 Missing data imputation via K-Nearest Neighbors

In many real-world datasets—particularly those derived from large-scale international repositories such as the OECD—missing data represent a critical challenge for quantitative analysis. In this study, the OECD regional indicators dataset included various missing entries, denoted in the original files by placeholder values such as `".."`. These gaps, if left unaddressed, could compromise the validity of statistical models and bias the results, especially when multivariate techniques are applied.

To ensure data completeness while preserving the structural relationships among variables, we adopted a non-parametric imputation approach based on K-Nearest Neighbors (KNN), implemented via the `KNNImputer` module in `scikit-learn`. Unlike simple mean or median imputation, KNN leverages local similarity among observations to estimate plausible values, making it particularly suitable for multidimensional datasets where correlations among features are strong and meaningful.

The procedure was executed in a batch-processing loop across multiple CSV files. For each file, the process followed four key steps. First, the dataset was imported, and all placeholder strings (specifically `".."`) were converted to `NaN` using the `na_values` parameter in `pandas.read_csv`. This step ensured consistent handling of missing data across all datasets.

Second, the script separated numeric variables from categorical ones using the `select_dtypes` function, restricting the imputation procedure exclusively to quantitative features. This is essential because KNN-based imputation relies on Euclidean distance calculations, which are not defined for non-numeric data. As such, all non-numeric columns (e.g., region names, country codes) were left untouched.

Third, the imputation logic explicitly excluded the target variables—namely `Life satisfaction` and `Self assessment of life satisfaction`—from the set of features used in the KNN model. This decision reflects a deliberate design choice: since these variables serve as dependent or outcome variables in subsequent modeling stages, imputing them could introduce artificial structure into the data or confound the interpretation of prediction results. Consequently, only explanatory variables were subjected to KNN-based imputation.

The actual imputation was performed using five nearest neighbors (`n_neighbors=5`). For each missing value in a feature, the algorithm identified the five most similar instances (based on non-missing values across the feature space) and computed the mean of their

corresponding values in the target column. This estimated value was then assigned to the missing entry. The imputation was conducted in a column-wise manner, preserving the internal coherence of the dataset and minimizing the risk of distortion.

Finally, the imputed datasets were saved in-place, overwriting the original CSV files. This ensured seamless continuity with subsequent analytical steps while maintaining version control at the file level.

The use of KNN for imputation offers several advantages in this context. It adapts flexibly to local patterns in the data without imposing rigid parametric assumptions. Moreover, it handles complex, non-linear relationships between variables, which is particularly relevant in the OECD well-being framework, where indicators span diverse domains such as income, education, environment, and health. By applying KNN imputation exclusively to the explanatory variables and preserving the integrity of the life satisfaction metrics, we ensured both analytical rigor and model interpretability in all subsequent steps of the project.

Chapter 3

Model definition

3.1 Data extraction and structuring from OECD sources

3.1.1 Overview of source files

The original data for this project was retrieved from the official OECD Regional Well-Being platform in the form of a multi-sheet Excel file (`OECD_all.xlsx`). This spreadsheet includes various datasets organized across different tabs, corresponding to different conceptualizations and timeframes of well-being indicators—such as current scores, raw numerical indicators, and longitudinal trends. To extract these datasets in a structured and reproducible manner, we designed a modular Python routine using `pandas` and `openpyxl`.

The first step involved inspecting the available sheet names within the Excel file using `pandas.ExcelFile`, which confirmed the presence of multiple relevant sheets, including `Score_Last`, `Indicator_Last`, and `Score_Trend`. These sheets provide, respectively, the normalized well-being scores at the regional level, raw structural indicators, and temporal trends for each region. For the purposes of this study, we focused exclusively on the regional-level scores and indicators, discarding trend data and national-level summaries in the modeling phase.

3.1.2 Extraction of score data (regional and national)

The tab `Score_Last` contains standardized scores for 11 well-being dimensions across 447 OECD subnational regions. These scores are normalized on a 0–10 scale and represent the relative positioning of each region within the OECD distribution.

To extract this information, we defined a dedicated function `load_score_last` specifying both the sheet name and the column range (`B:O`) to be loaded. The data was imported using row 6 as the header (to capture official variable names), with row 7 skipped due to formatting artifacts. A total of 447 rows were imported—corresponding to all available regions—and the resulting DataFrame was saved as `Score_Last_Region.csv` for further analysis.

Additionally, a national-level summary table was extracted from rows 458 to 496 of the same sheet. These rows contain aggregate values for each country, matching the structure of the regional dataset. We assigned consistent column names by copying those from the regional dataset and removing the second column, which is blank or redundant. This subset was saved as `Score_Last_Country.csv`, allowing for high-level comparisons across countries when needed.

3.1.3 Extraction of structural indicators (regional and national)

The `Indicator_Last` sheet contains unnormalized, raw values for structural well-being dimensions—such as unemployment rates, income levels, pollution concentrations, and broadband access. These indicators provide the empirical backbone for later modeling and clustering.

Using a similar function (`load_indicator_last`), we imported columns `B:R` from the `Indicator_Last` sheet, again specifying row 6 as the header and skipping row 7. Each column corresponds to a different indicator, often measured in heterogeneous units (e.g., USD, percent, $\mu g/m^3$). To make the variables interpretable, a suffix string was parsed and programmatically appended to each column name, explicitly annotating the unit of measurement (e.g., "per 100,000 people" for homicide rates or "$\mu g/m^3$" for PM2.5 pollution). This labeling process ensures semantic clarity during downstream analyses.

As with the score data, a national summary table was also extracted from rows 458 to 496, and the columns were harmonized with those in the regional file. This table was saved as `Indicator_Last_Country.csv`, although it was not used directly in modeling.

3.1.4 Extraction of trend data (not used in analysis)

Although not used in the main analyses, we also extracted the `Score_Trend` sheet, which contains longitudinal information on well-being scores. The structure is similar to that of `Score_Last`, covering 447 regions. This dataset was saved as `Score_Trend_Region.csv` for potential future exploration of temporal dynamics, although it was excluded from the modeling pipeline presented in this report.

3.1.5 Final dataset inventory and usage scope

In total, five datasets were extracted and saved as separate CSV files:

- `Score_Last_Region.csv`: standardized well-being scores at the regional level.
- `Score_Last_Country.csv`: country-level summary of the same scores.
- `Indicator_Last_Region.csv`: raw structural indicators at the regional level.
- `Indicator_Last_Country.csv`: country-level summary of indicators.
- `Score_Trend_Region.csv`: regional well-being scores over time.

For all analyses described in this project—including modeling, clustering, and visualization—we exclusively employed the regional versions of the score and indicator datasets. These datasets offer the most granular and policy-relevant information for studying subnational disparities in well-being and constructing scientifically grounded regional typologies.

3.2 Missing data imputation and target isolation

3.2.1 KNN imputation strategy

After extracting the raw datasets, a crucial preprocessing step involved handling missing data. In both the `Score_Last_Region.csv` and `Indicator_Last_Region.csv` files, missing values are represented by the placeholder string `".."`. To ensure consistency and

compatibility with analytical models, these entries were first converted to standard `NaN` values upon import using the `na_values=[".."]` argument within `pandas.read_csv`.

Given the multidimensional nature of the data and the importance of preserving inter-variable relationships, we employed K-Nearest Neighbors (KNN) imputation. This non-parametric technique estimates missing values by computing the average of the $k = 5$ most similar instances (in terms of Euclidean distance) based on the observed values across all other numeric features. KNN is particularly suited for datasets like this one, where multiple dimensions are moderately correlated, and imputing with mean or median would oversimplify underlying dependencies.

Importantly, variables corresponding to the modeling target—specifically `Life satisfaction` and `Self assessment of life satisfaction`—were explicitly excluded from the imputation process. This decision ensures that the predictive task remains unbiased and not artificially influenced by values inferred from correlated features. By excluding the target variables from the feature set passed to `KNNImputer`, we prevent data leakage and ensure the reliability of the subsequent supervised learning models.

Once the imputation was completed for the explanatory variables, the cleaned datasets were saved, overwriting the originals, thereby preserving a uniform data format for all future analyses.

3.2.2 Dataset import and target cleansing

Following the imputation procedure, the regional score dataset (`Score_Last_Region.csv`) was re-imported and prepared for predictive modeling. The first operation consisted in the removal of rows with missing values in the dependent variable `Life satisfaction`, which serves as the primary outcome of interest for regression models. This step led to the exclusion of 10 observations out of 447, ensuring that all training samples include valid target values.

Subsequently, a clean feature matrix X was created by dropping all non-numeric and non-predictive columns: specifically, the geographic identifiers `Country`, `Region`, and `Code`, along with the target itself. The resulting dataset contains only explanatory features, all of which are numerical and imputed, ready to be used in supervised regression models. The response vector y was defined as the `Life satisfaction` column, thus formalizing the supervised learning task.

This two-stage process—KNN-based imputation followed by exclusion of incomplete target values—ensures both completeness of the predictor space and integrity of the modeling objective. It establishes a clean and coherent modeling dataset, minimizing noise and bias due to missing data and laying the foundation for robust predictive inference.

3.3 Modeling pipeline: train-test split and hyperparameter optimization on the Score dataset

3.3.1 Train-test split procedure

To enable a robust and unbiased evaluation of predictive models, we partitioned the dataset into training and testing subsets using an 80/20 split. The operation was performed via the `train_test_split` function from the `scikit-learn` library, with a fixed random seed (`random_state=1748`) to ensure reproducibility. This split guarantees that

the training set contains a representative sample of regional units for model fitting, while the test set provides an independent benchmark for out-of-sample performance assessment. The indices corresponding to the test observations were stored for later use in evaluation and visualization steps.

3.3.2 Hyperparameter optimization: coarse-to-fine tuning

In order to maximize the predictive accuracy and generalization ability of our models, we implemented a two-stage hyperparameter optimization process based on exhaustive grid search. The candidate models were **Random Forest** and **XGBoost**, both of which are ensemble-based tree models well-suited for tabular regression tasks with complex interactions among predictors.

Stage I: coarse search

The first tuning phase was exploratory in nature and involved relatively broad hyperparameter grids. For Random Forest, we varied the number of trees (`n_estimators`) between 100 and 500, evaluated several values for tree depth (`max_depth`), and tested different configurations for the minimum number of samples required to split internal nodes (`min_samples_split`) and to constitute a leaf (`min_samples_leaf`), as well as both bootstrapped and non-bootstrapped strategies. For XGBoost, the search space encompassed learning rates from 0.01 to 0.1, tree depths from 3 to 6, and a range of values for regularization parameters such as `gamma`, `subsample`, and `colsample_bytree`. Each model was tuned via 5-fold cross-validation using negative root mean squared error (RMSE) as the scoring metric.

The best performing configurations identified were:

- Random Forest: `n_estimators=100`, `max_depth=None`, `min_samples_split=2`, `min_samples_leaf=` `bootstrap=True`, with a cross-validated RMSE of 1.454.
- XGBoost: `n_estimators=300`, `max_depth=6`, `learning_rate=0.05`, `subsample=0.5`, `colsample_bytree=0.8`, `gamma=0.1`, achieving a superior cross-validated RMSE of 1.431.

On the held-out test set, XGBoost outperformed Random Forest with an RMSE of 1.280 versus 1.354.

Stage II: fine-grained tuning

Building upon the initial results, a second round of tuning was conducted using narrower intervals centered around the optimal values from the first phase. For Random Forest, we explored a refined range of `n_estimators` (from 90 to 110), focused on `min_samples_split` values near 2 and 3, and restricted the search to bootstrapped models with default maximum depth. For XGBoost, the grid was similarly constrained around the best preliminary values for all six hyperparameters.

The fine-tuned results closely confirmed those obtained in the initial round:

- Random Forest: best configuration was `n_estimators=110`, `min_samples_split=3`, with a cross-validated RMSE of 1.451 and a test RMSE of 1.353.

- XGBoost: retained its optimal configuration with `n_estimators=305`, `max_depth=6`, `learning_rate=0.05`, `subsample=0.5`, `colsample_bytree=0.8`, and `gamma=0.1`, achieving a cross-validated RMSE of 1.431 and a test RMSE of 1.280.

These results reinforce the robustness of the XGBoost model under both coarse and fine-grained tuning regimes and support its selection as the primary model for downstream interpretation and policy simulation. Random Forest also demonstrates strong performance and remains a valid benchmark for comparative analysis.

3.3.3 Model comparison and final selection

To determine the most suitable predictive model for explaining regional variations in subjective well-being (*Life satisfaction*), we compared the performance of the fine-tuned Random Forest and XGBoost regressors across multiple evaluation metrics and statistical tests.

Both models were trained using the best hyperparameters obtained from the fine-tuning stage, and their performance was assessed using 5-fold cross-validation with three standard metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). For each metric, we also conducted a paired t-test to assess whether observed differences in performance were statistically significant.

Root Mean Squared Error (RMSE). Random Forest achieved a lower cross-validated RMSE (mean = 1.4653, std = 0.0693) compared to XGBoost (mean = 1.5514, std = 0.0998). The paired t-test yielded a t-statistic of 4.1648 and a p-value of 0.0141, indicating a statistically significant difference in favor of Random Forest at the 5% significance level.

Mean Absolute Error (MAE). Similarly, Random Forest outperformed XGBoost in terms of MAE (mean = 1.0963, std = 0.0726 vs. mean = 1.1743, std = 0.0509). The corresponding paired t-test reported a t-statistic of 5.8103 and a p-value of 0.0044, again confirming the superiority of Random Forest with statistical significance.

Coefficient of determination (R^2). With regard to R^2, Random Forest achieved a higher average score (mean = 0.7004, std = 0.0473) than XGBoost (mean = 0.6636, std = 0.0586). The t-test resulted in a t-statistic of -4.0960 and a p-value of 0.0149, supporting the conclusion that the Random Forest model exhibits a statistically significant higher explanatory power.

Conclusion. Across all three performance metrics—RMSE, MAE, and R^2—Random Forest consistently outperformed XGBoost. The results of the paired t-tests further confirm that these differences are not due to random fluctuations but represent a true performance advantage. As a result, the Random Forest regressor was selected as the final model for interpretation, explainability analysis, and policy simulation tasks in the subsequent sections of this report.

3.4 Prediction diagnostics and residual analysis

Following the selection of the Random Forest model as the optimal regressor, we performed a visual and statistical evaluation of its predictive behavior. This step serves to assess the goodness-of-fit, identify potential model misspecifications, and detect systematic deviations or outliers that might compromise interpretability.

3.4.1 Predicted vs actual values

The scatter plot comparing predicted values against observed life satisfaction scores (Figure 3.1) shows a strong linear trend, with most points concentrated along the 45-degree reference line. This alignment indicates a satisfactory level of agreement between predicted and actual values. However, some degree of dispersion is visible, particularly for observations at the lower and upper ends of the scale. The overall spread confirms that the model captures a substantial portion of the signal but also exhibits some prediction variability, especially in extreme cases.

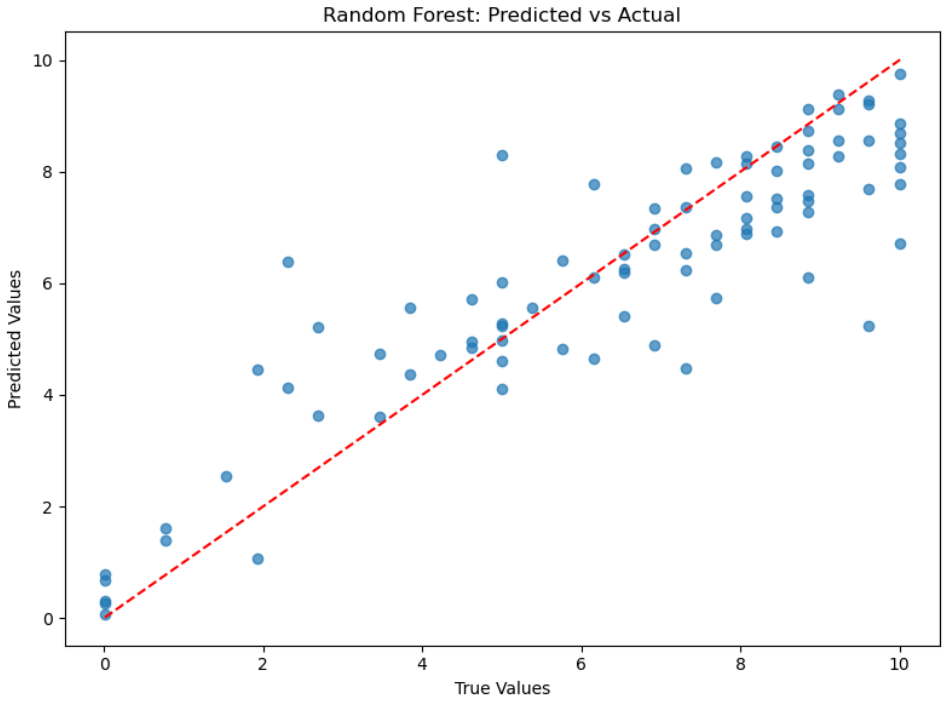


Figure 3.1: Predicted vs Fitted values

3.4.2 Residual plot

The residual plot (Figure 3.2) displays residuals as a function of predicted values. Ideally, the residuals should appear randomly scattered around the zero line, without discernible patterns. This condition largely holds in our case, although some heteroskedasticity can be observed: the variance of the residuals increases slightly for predicted values above 7. This behavior may indicate that the model is slightly less precise in regions with higher life satisfaction, possibly due to greater heterogeneity in well-being determinants at those levels.

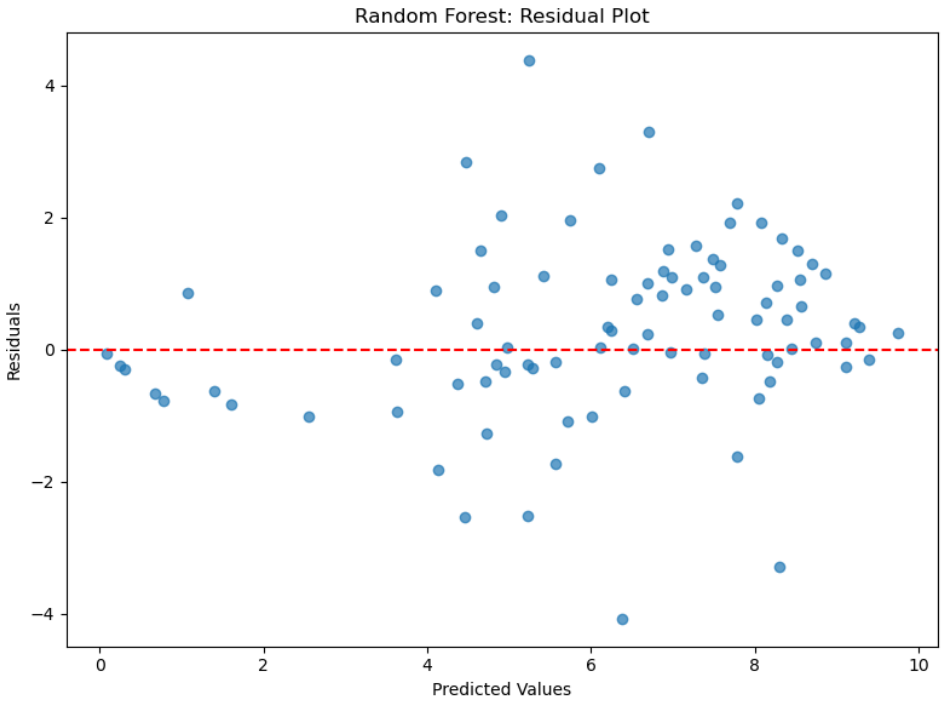


Figure 3.2: Residual plot

3.4.3 Residual Distribution and Summary Statistics

The histogram of residuals (Figure 3.3) approximates a bell-shaped curve centered near zero, indicating that the prediction errors are roughly symmetrically distributed. While not perfectly Gaussian, the residual distribution does not show strong skewness or kurtosis, and its visual appearance supports the assumption of error balance.

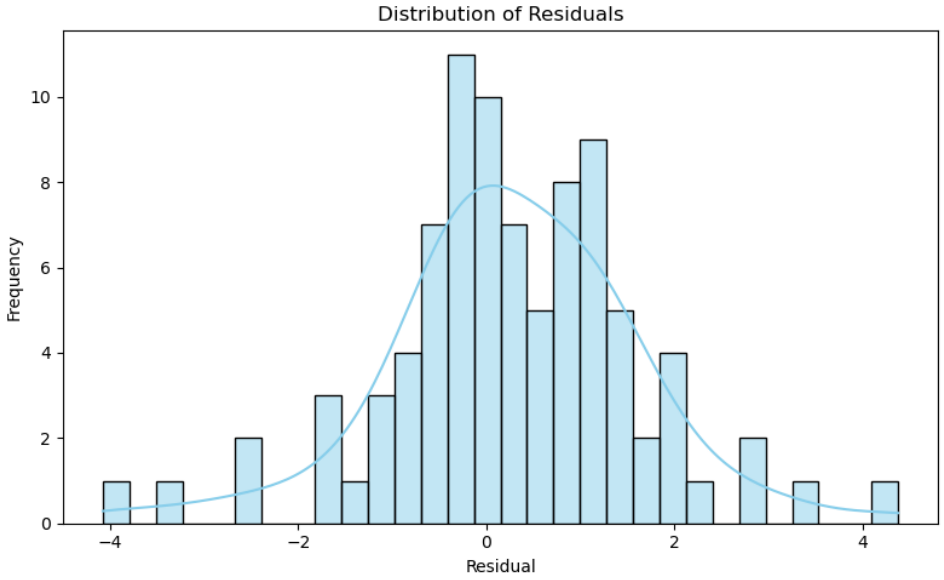


Figure 3.3: Residual distribution

The summary statistics further confirm this observation. The mean residual is slightly

positive (0.28), the median is close to zero (0.25), and the interquartile range lies between -0.36 and 1.07. This suggests that the majority of predictions fall within an acceptable error margin. However, the range spans from -4.08 to +4.37, highlighting the presence of some notable outliers. These may correspond to regions whose well-being profiles are highly atypical or influenced by unobserved contextual factors not captured in the feature set.

3.4.4 Interpretation

Overall, the residual analysis confirms the robustness and reliability of the Random Forest model in predicting regional life satisfaction. The presence of a few outliers is not uncommon in large-scale socioeconomic datasets and may merit further case-specific investigation. The model maintains unbiased predictions on average, with no major systematic deviation or violation of core regression assumptions.

3.5 SHAP-based interpretability analysis

3.5.1 Global feature importance via Mean Absolute SHAP values

To interpret the internal decision logic of the Random Forest model and quantify the contribution of each predictor to the model's output, we employed SHAP (SHapley Additive exPlanations), a game-theoretic approach to explain individual predictions. We began by computing the mean absolute SHAP values across all training samples, thereby obtaining a global feature importance ranking (Figure 3.4).

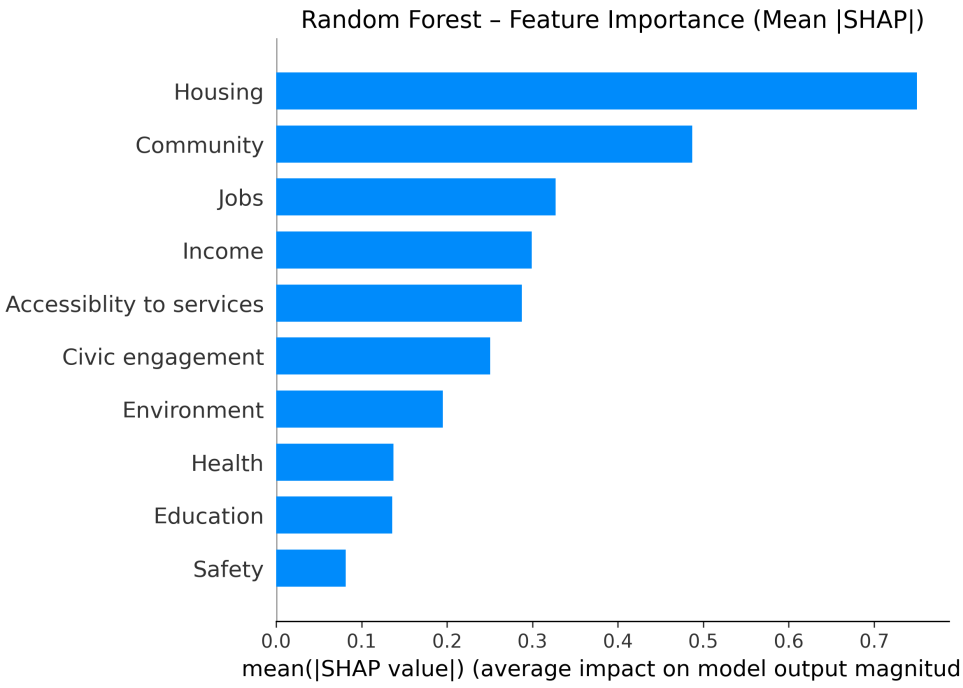


Figure 3.4: SHAP summary bar plot

Among all features, **Housing** emerges as the most influential, with a mean SHAP

value of approximately 0.75. This indicates that variation in housing-related indicators systematically contributes the most to fluctuations in predicted life satisfaction. **Community**, **Jobs**, and **Income** also show high explanatory power, reinforcing the relevance of both social cohesion and economic opportunity in determining subjective well-being. In contrast, **Safety**, **Education**, and **Health** exhibit lower average contributions, suggesting either weaker associations or potential nonlinear patterns that reduce their overall influence.

3.5.2 Distributional effects: SHAP Beeswarm plot

To better understand the distributional impact of features across individual predictions, we visualized a SHAP beeswarm plot (Figure 3.5). Each dot represents a region; its horizontal position reflects the SHAP value (i.e., contribution to the prediction), while its color encodes the feature's value (from low in blue to high in red).

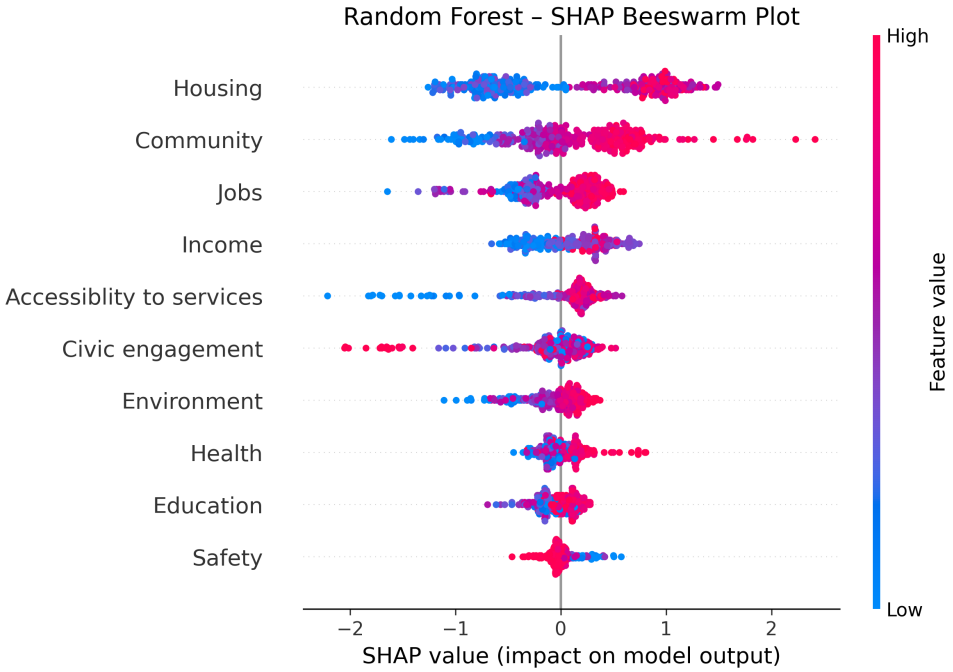


Figure 3.5: Random Forest – SHAP Beeswarm Plot

Most features exhibit expected monotonic behavior—higher input values lead to increased predictions. However, certain features present non-trivial or even counterintuitive patterns. For instance, while higher **Housing** values consistently raise predicted satisfaction, **Safety** shows an inverted association: higher safety scores often reduce predicted outcomes. This could point to collinearity, measurement biases, or underlying sociopolitical complexities.

Similarly, **Civic Engagement** displays dual behavior—while generally contributing positively, some high values are linked to lower predictions, hinting at potential subgroup heterogeneity or nonlinear interactions. Features such as **Accessibility to Services**, **Environment**, and **Health** cluster around zero SHAP impact for most regions, with larger effects appearing primarily at low values. This suggests that deficiencies in these domains may depress well-being more than improvements enhance it, consistent with a loss-aversion dynamic.

3.5.3 Feature-wise SHAP dependence patterns

To further explore feature-behavior relationships, we generated SHAP dependence plots for the top 10 predictors (Figure 3.6), mapping each feature's value against its SHAP contribution.

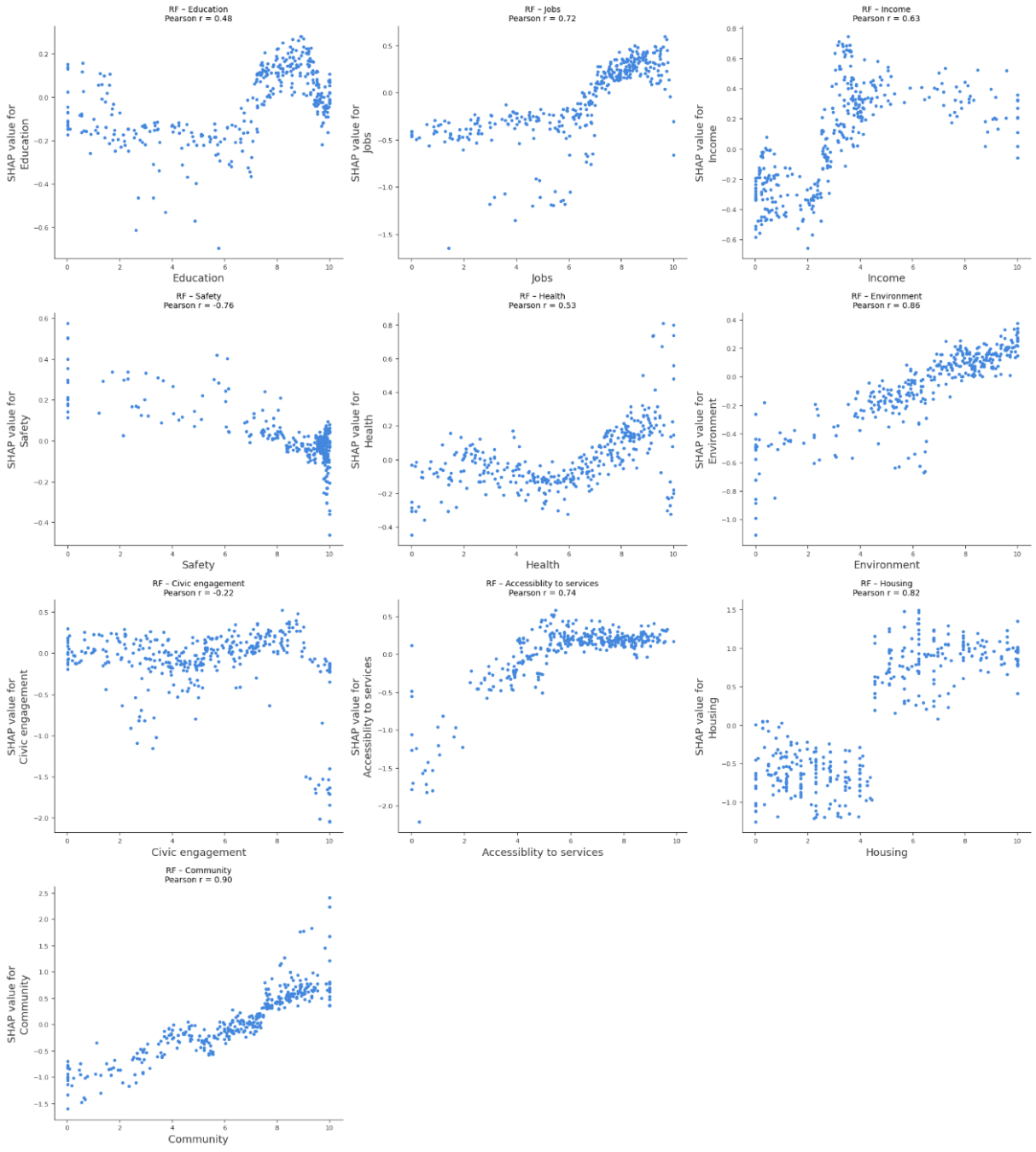


Figure 3.6: SHAP dependence matrix

Several distinct functional forms emerge:

- **Jobs** and **Community** show strong, mostly linear positive associations, with higher values yielding greater predicted satisfaction.
- **Income** rises steeply with SHAP impact up to a threshold (around score 4), then plateaus, indicating diminishing returns.

- **Housing** exhibits a bimodal pattern—scores below 5 correlate with negative SHAP values, while scores above 5 show a modest positive effect.
- **Safety** demonstrates a surprising inverse correlation, with a sharp drop in SHAP value for scores near 10.
- **Education** and **Health** suggest U-shaped patterns, where both very low and very high values contribute positively, while intermediate values offer little predictive weight.

These insights align with psychological theories that emphasize both absolute and relative dimensions of well-being. For example, **Community** consistently emerges as a key enhancer of life satisfaction, echoing findings from well-being literature on social capital. Conversely, the weak or non-monotonic impact of **Education** and **Safety** may reflect their indirect or mediated role in subjective well-being.

3.5.4 Interpretative summary

SHAP analysis provides a robust, transparent decomposition of model predictions, highlighting both global and local mechanisms. It confirms the high predictive relevance of structural indicators such as housing quality and social connectedness, while revealing nuanced behaviors for variables traditionally assumed to be unilaterally beneficial (e.g., safety or education). This multidimensional view is critical for designing evidence-based, targeted well-being policies that account for both synergies and trade-offs across domains.

3.6 Modeling pipeline: train-test split and hyperparameter optimization on the Indicators dataset

3.6.1 Train-test split procedure

To construct predictive models on the Indicators dataset, we first removed all regional entries lacking valid values for the target variable, *Self assessment of life satisfaction*. This preprocessing step excluded 10 observations and ensured that the target remained well-defined across all rows.

We then partitioned the dataset into training and test subsets using an 80/20 split. The operation was conducted via the `train_test_split` utility from `scikit-learn`, with the `random_state` parameter fixed at 1748 to maintain reproducibility. This stratified split guarantees a robust estimate of generalization performance, as well as a fair comparison between models. Test indices were stored for subsequent evaluation and visualization.

3.6.2 Hyperparameter optimization: coarse-to-fine tuning

To optimize the predictive performance of our models on the `Indicators` dataset, we adopted a two-stage hyperparameter tuning strategy for both Random Forest and XGBoost regressors. The tuning process was conducted via 5-fold cross-validation using the negative Root Mean Squared Error (RMSE) as the scoring metric. All experiments were performed with a fixed random seed (`random_state = 1748`) to ensure reproducibility.

Stage I: coarse grid search

In the initial stage, we performed an extensive grid search to explore a wide range of hyperparameter combinations. For the Random Forest model, we varied parameters such as the number of trees (`n_estimators`), tree depth, and minimum split and leaf sizes. For the XGBoost regressor, the grid included learning rate, depth, column and row sampling ratios, and regularization strength.

The best configurations identified through this coarse search were:

- **Random Forest**: `n_estimators=200`, `max_depth=None`, `min_samples_split=2`, `min_samples_leaf=1`, `bootstrap=True`
- **XGBoost**: `n_estimators=500`, `max_depth=3`, `learning_rate=0.05`, `subsample=1.0`, `colsample_bytree=0.7`, `gamma=0`

The corresponding cross-validated RMSE scores were:

- Random Forest: 0.413
- XGBoost: 0.407

When tested on the hold-out test set, both models showed excellent generalization performance:

- Random Forest Test RMSE: 0.383
- XGBoost Test RMSE: 0.385

Stage II: fine grid search

After identifying the most promising regions of the hyperparameter space, we refined our search using a narrower grid centered around the optimal values from the coarse phase. This stage was designed to stabilize performance and capture marginal gains in predictive accuracy.

The refined configurations returned by the fine grid search were:

- **Random Forest**: `n_estimators=205`, `max_depth=None`, `min_samples_split=2`, `min_samples_leaf=1`, `bootstrap=True`
- **XGBoost**: `n_estimators=505`, `max_depth=3`, `learning_rate=0.05`, `subsample=1.0`, `colsample_bytree=0.7`, `gamma=0`

The validation scores remained consistent with the earlier stage:

- Random Forest CV RMSE: 0.413
- XGBoost CV RMSE: 0.407

Final evaluation on the test set confirmed the robustness of both models:

- Random Forest Test RMSE: 0.383
- XGBoost Test RMSE: 0.384

These results demonstrate that both models perform competitively on the `Indicators` dataset, with Random Forest maintaining a slight edge in test performance.

3.6.3 Model comparison and final selection

To assess which of the two candidate models—Random Forest and XGBoost—offers superior performance on the `Indicators` dataset, we conducted a model comparison based on three standard regression metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2). All comparisons were made via 5-fold cross-validation, followed by paired t-tests to evaluate the statistical significance of observed differences.

Cross-validation results: RMSE

We first evaluated RMSE scores across five folds for both models. The results were as follows:

- **XGBoost**: Mean RMSE = 0.4479, Standard Deviation = 0.0561
- **Random Forest**: Mean RMSE = 0.4391, Standard Deviation = 0.0650

A paired t-test on RMSE scores yielded a t-statistic of 1.2246 and a p-value of 0.2879. As the p-value exceeds the standard 0.05 threshold, the difference between the models is not statistically significant with respect to RMSE.

Cross-validation results: MAE

For MAE, Random Forest achieved a slightly lower error:

- **XGBoost**: Mean MAE = 0.3196, Standard Deviation = 0.0190
- **Random Forest**: Mean MAE = 0.3004, Standard Deviation = 0.0162

Here, the paired t-test indicated statistical significance, with a t-statistic of 3.1282 and a p-value of 0.0352. This result favors Random Forest as the more accurate model in terms of absolute prediction error.

Cross-validation results: R^2

Lastly, model performance was evaluated using the R^2 coefficient:

- **XGBoost**: Mean R^2 = 0.6461, Standard Deviation = 0.0459
- **Random Forest**: Mean R^2 = 0.6585, Standard Deviation = 0.0645

The paired t-test for R^2 differences returned a t-statistic of -1.1544 and a p-value of 0.3126, indicating that the performance difference is not statistically significant on this metric either.

Model selection

Among the three performance metrics considered, only the difference in MAE was found to be statistically significant, in favor of Random Forest. While both models performed comparably in terms of RMSE and R^2, the slightly superior performance and lower variability of Random Forest on the MAE metric led us to select it as the final model for further interpretation and SHAP analysis.

Additionally, considering that Random Forest also outperformed XGBoost on the `Score` dataset, its selection enables a more coherent comparison of feature influence across both modeling pipelines.

3.7 Prediction diagnostics and residual analysis

3.7.1 Predicted vs actual and residual plot

To assess the predictive performance of the Random Forest model trained on the *Indicators* dataset, we examined the relationship between predicted and observed values. The first scatter plot presents this comparison, where the dashed red line represents the ideal scenario of perfect predictions. The majority of points lie close to this line, indicating a strong alignment between the model's outputs and the actual values, particularly in the mid-range of the target variable.

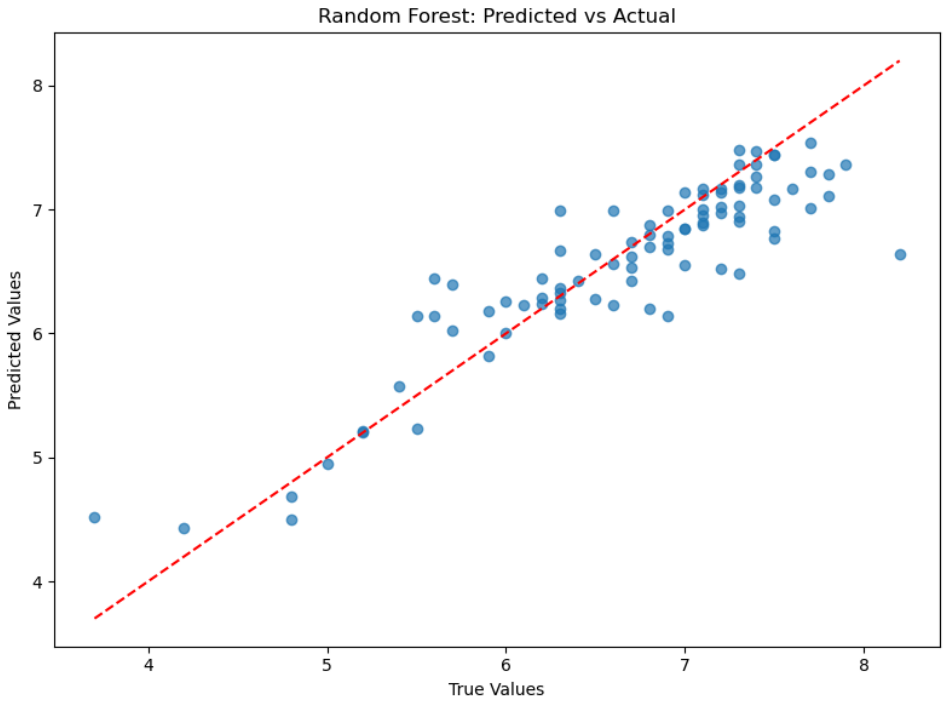


Figure 3.7: Predicted vs actual values (Indicators dataset)

The second plot illustrates the residuals plotted against the predicted values. The residuals are symmetrically distributed around zero and show no discernible structure or trend, suggesting that the model does not suffer from heteroskedasticity or systematic bias. This supports the assumption of independently and identically distributed errors and confirms that the model generalizes well across the prediction spectrum.

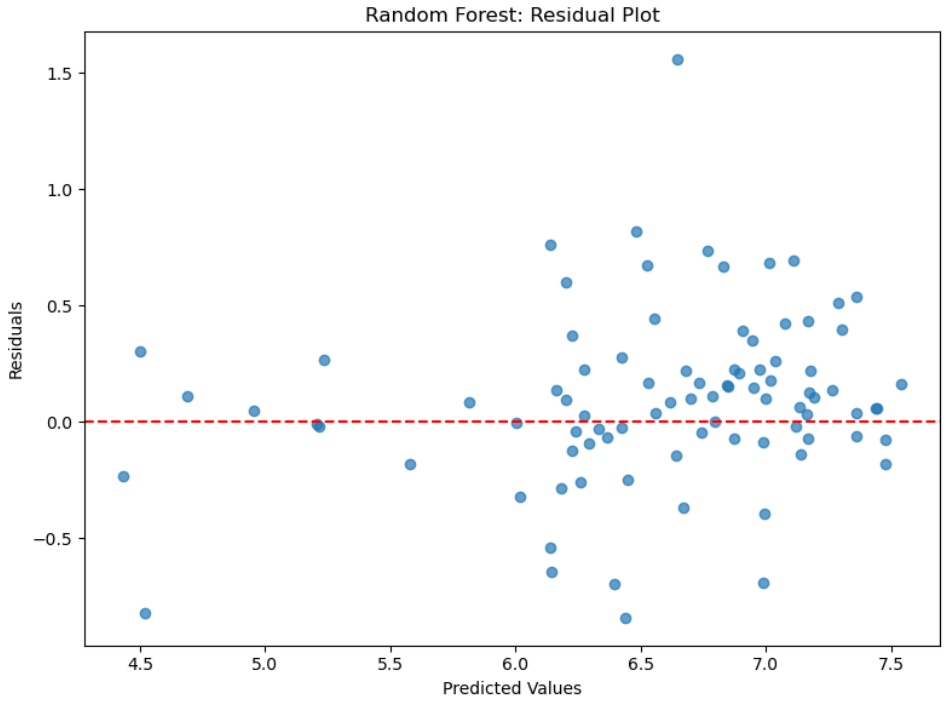


Figure 3.8: Residual plot (Indicators dataset)

3.7.2 Residual distribution and summary statistics

The third plot presents the distribution of residuals, complemented by a kernel density estimate. The bell-shaped form approximates a normal distribution with a slight positive skew. Summary statistics reveal a residual mean of 0.096 and a standard deviation of 0.373, indicating that the model's predictions are, on average, unbiased and moderately dispersed. The residuals range from approximately -0.84 to +1.56, implying that while most predictions are accurate, a few exhibit larger deviations.

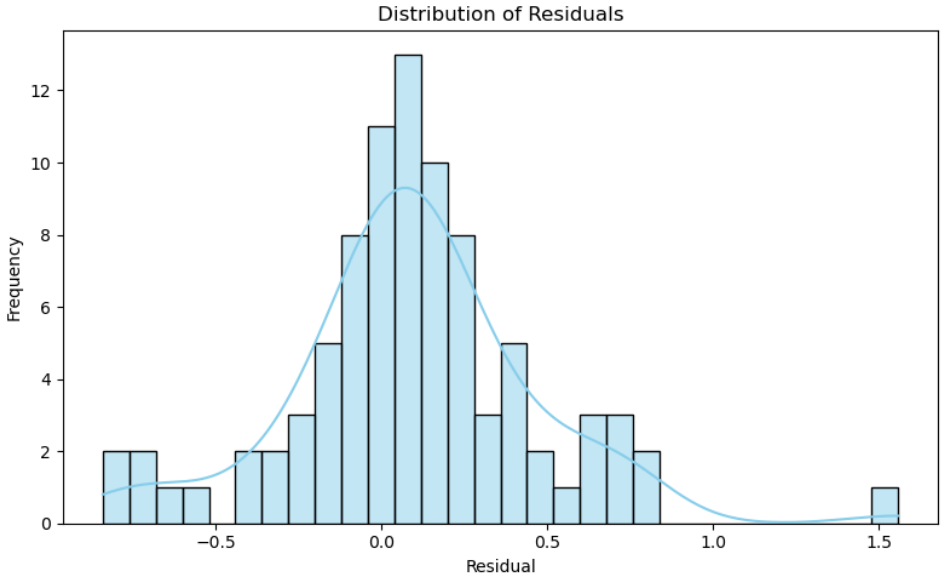


Figure 3.9: Distribution of the residuals (Indicators dataset)

Overall, the Random Forest model demonstrates strong predictive performance on the *Indicators* dataset, with residuals exhibiting lower magnitude and variability compared to those obtained from the *Score* dataset. This suggests greater consistency and accuracy, thereby reinforcing the model's suitability for interpretability analysis through SHAP methods.

3.8 SHAP analysis

3.8.1 SHAP table and bar chart

The SHAP analysis was conducted on the Random Forest model trained with the `indicators` dataset. The bar plot of mean absolute SHAP values highlights the relative importance of each predictor. The most influential features are *Perceived social network support*, *Number of rooms per person*, *Household disposable income per capita*, and *Employment rate*, each contributing significantly to the model's prediction of self-reported life satisfaction.

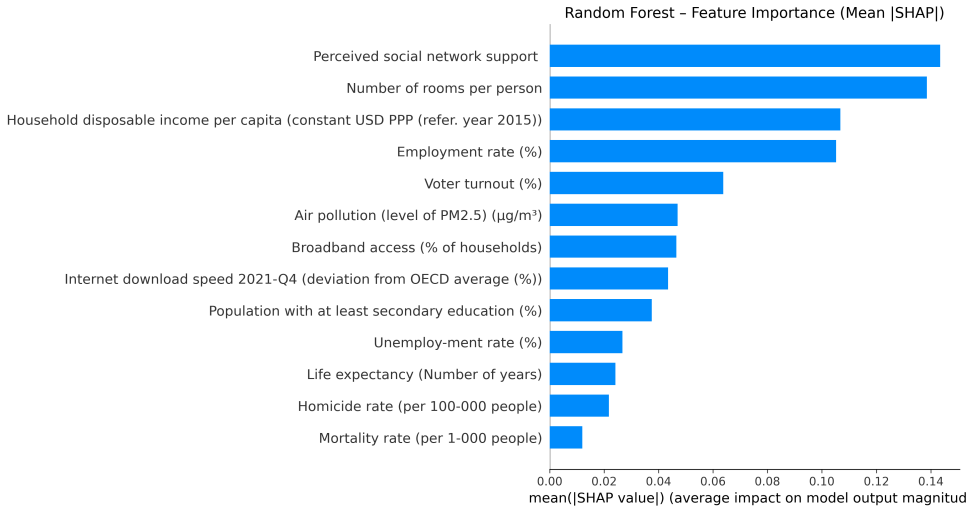


Figure 3.10: Feature importance in the Indicators dataset

Although the SHAP values obtained are notably smaller than those derived from the `score` dataset, this difference is likely attributable to the narrower range of the target variable in the indicators data. Importantly, what matters is not the absolute value of SHAP scores but their relative magnitude, which reflects how much each feature influences the model's predictions.

A comparison of SHAP-based rankings across both models shows a strong alignment in the importance of broad dimensions such as Housing, Community, Jobs, and Income. Minor discrepancies—such as lower ranks for Accessibility and Health indicators—may be driven by the aggregation level of variables and their measurement scale.

3.8.2 SHAP beeswarm plot

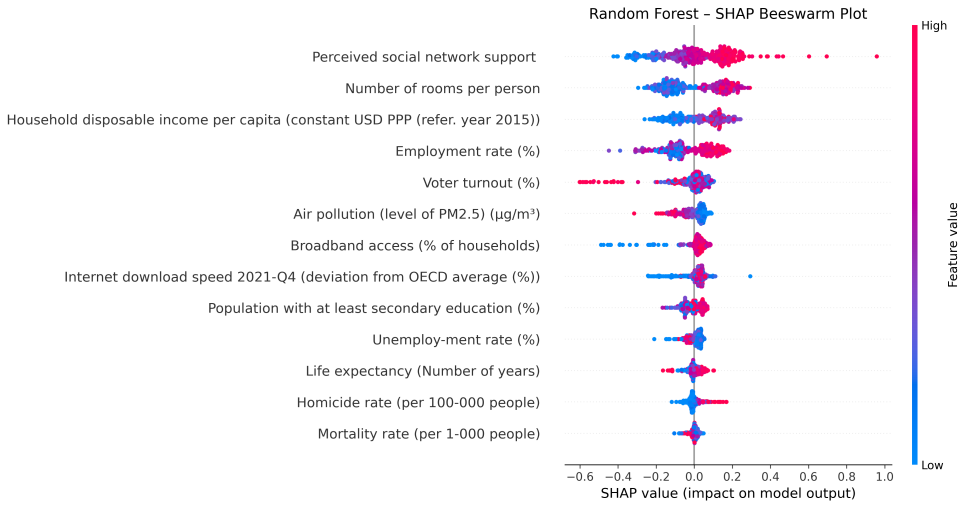


Figure 3.11: SHAP beeswarm plot for Indicators dataset

The SHAP beeswarm plot further illustrates the distribution of feature effects. The feature *Number of rooms per person* behaves analogously to the Housing score, while *Perceived social network support* mirrors the pattern of the Community score. Similarly,

Household disposable income per capita reflects the dynamics of the Income score. For Accessibility, *Broadband access* emerges as a stronger contributor than *Internet download speed*, consistent with the interpretation that access represents a prerequisite, whereas speed is a secondary factor. For employment-related indicators, *Employment rate* and *Unemployment rate* show complementary patterns: higher employment values (in red) are associated with positive SHAP values, while higher unemployment values (also in red) correspond to negative contributions. These opposing trends are consistent with the composite behavior observed in the Jobs score.

In the case of *Voter turnout*, the distribution is more centered but still preserves a directionality similar to the Civic Engagement score. A similar alignment is observed between *Population with at least secondary education* and the Education score, as well as between *Air pollution* and Environment. The features *Homicide rate* and *Mortality rate* exhibit expected inverse contributions compared to *Safety* and *Life expectancy*, respectively.

3.8.3 SHAP dependence plots

The dependence plots reveal detailed relationships between individual predictors and their SHAP values. Several variables exhibit non-linear or threshold effects.

Perceived social network support shows a strong positive linear correlation with SHAP values ($r = 0.71$), paralleling the Community dimension. *Number of rooms per person* displays a cluster around mid-range values with moderate dispersion, consistent with its housing-related interpretation.

Household disposable income presents a highly non-linear pattern, with a rapid increase in SHAP contribution after a certain income threshold. *Employment rate* shows a clearly positive effect after surpassing 70%, whereas *Unemployment rate* confirms a generally negative influence, albeit with notable outliers even at low values.

Air pollution is strongly negatively correlated with SHAP values ($r = -0.69$), while *Broadband access* is positively correlated, with most data concentrated above 70%. *Internet download speed* exhibits a saturation effect, with SHAP values flattening after reaching a deviation threshold. These patterns are consistent with the trends identified in the score-based Accessibility variable.

The education-related feature shows a U-shaped pattern, where extreme values (both low and high) correspond to more positive SHAP contributions, replicating the dynamics seen in the Education score. *Voter turnout* and *Homicide rate* present directional impacts similar to their respective composite scores. Finally, the comparison of *Life expectancy* and *Mortality rate* reflects the expected inverse trends, although the separation is less distinct, echoing the mixed SHAP signal of the Health score in the first model.

Overall, the analysis highlights that the indicator-based model aligns well with the score-based model in terms of which dimensions drive life satisfaction. While the granularity of individual indicators allows for more nuanced interpretations—such as distinguishing between access and speed in digital infrastructure—it also confirms the robustness and validity of the dimension-level approach used in the previous analysis.

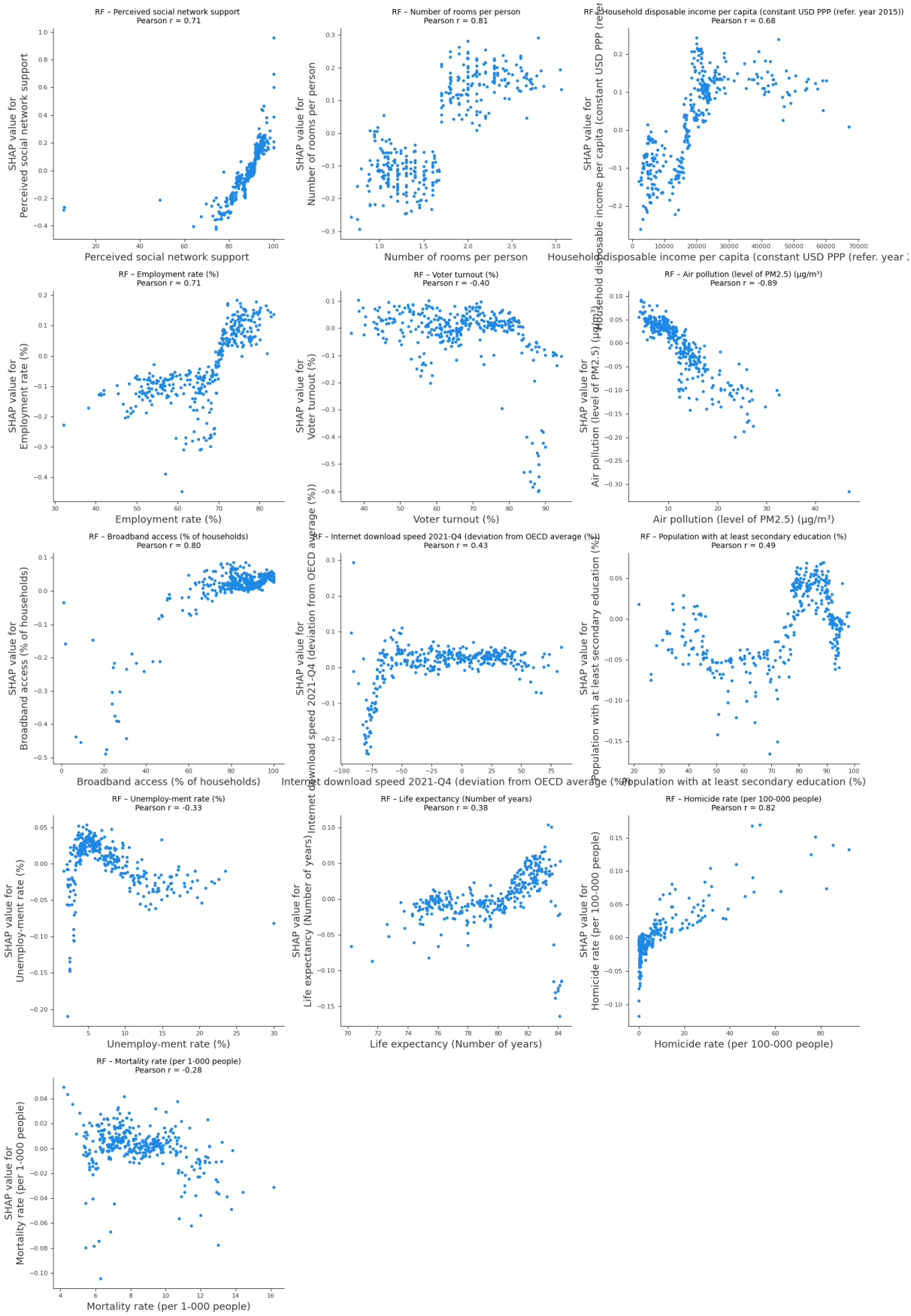


Figure 3.12: SHAP dependence matrix for Indicators dataset

3.9 Policy simulation

Policy simulations based on SHAP (SHapley Additive exPlanations) and Random Forest predictions allow for the interpretation of how changes in specific input features may affect life satisfaction outcomes. By manipulating individual variables in a median-profile observation, it is possible to quantify both the variation in the predicted score and the relative contribution of each feature. This approach offers a flexible and interpretable framework that can support data-driven policy evaluations.

Although the model should not be interpreted as a prescriptive tool for policymaking, it serves as a robust proof of concept. It highlights how machine learning combined with explainable AI techniques can guide policy priorities and assess the potential benefits of targeted interventions.

3.9.1 Baseline simulation - Score dataset

We first applied SHAP to a Random Forest model trained on the aggregated score dataset. The model prediction for the median case is 5.86. The SHAP decomposition shows that the variable with the largest negative contribution is **Housing** (-0.87), while **Income** ($+0.36$), **Accessibility to services** ($+0.29$), and **Jobs** ($+0.16$) emerge as positive drivers of life satisfaction. Minor but still relevant contributions come from **Environment**, **Civic Engagement**, and **Education**, whereas **Health** and **Safety** have limited impact.

Policy change simulation - Increasing Housing score by 30%

To simulate the effect of improving housing conditions, we increased the Housing score by 30%. This modification raised the value from 4.10 to 5.33 and resulted in a predicted life satisfaction of 7.46, an increase of 1.6 points. This is a substantial shift, largely driven by the reversal in the SHAP value for Housing, which changed from -0.87 to $+0.94$. Other variables such as **Income** ($+0.44$), **Education** ($+0.13$), and **Civic Engagement** (-0.14) showed moderate variation, confirming that their effects are influenced by feature interactions.

3.9.2 Baseline simulation - Indicator dataset

We then repeated the procedure on the indicator-based model. The prediction for the median case is 6.49. The largest negative contribution comes from the **Employment rate** (-0.12) and **Number of rooms per person** (-0.12), while **Income**, **Internet speed**, and **Broadband access** positively contribute to the score. As expected, variables such as **Air pollution**, **Unemployment**, and **Education** exert modest influence, while **Life expectancy**, **Mortality**, and **Homicide rate** are largely negligible.

Policy change simulation - Increasing Number of Rooms per Person by 30%

We simulated a housing-related intervention by increasing the number of rooms per person by 30%, from 1.62 to 2.11. The resulting prediction increased to 6.64. The SHAP value for this feature rose from -0.12 to $+0.14$, suggesting a considerable reversal in influence. Interestingly, this change also altered the SHAP values of other features, particularly **Employment rate** (from -0.12 to -0.20), **Education**, **Unemployment**, and **Income**,

which slightly increased in their contribution, reflecting the model's sensitivity to joint feature interactions.

3.9.3 Interpretation and insights

These simulations provide compelling evidence of the differential impact of policy domains depending on the granularity of data. When using aggregated scores, improvements in general housing conditions result in significant life satisfaction gains. The indicator-based model refines this finding, showing that a higher number of rooms per person is an effective proxy for housing quality and predicts higher well-being.

The simulations also reinforce the robustness of certain predictors across both models, notably income, housing, and access to services. Moreover, they reveal interaction effects not captured in traditional linear modeling. For example, the observed increase in income SHAP values after improving housing suggests a potential synergy between living conditions and economic capacity.

3.9.4 Conclusion

Policy simulations combining machine learning models with SHAP values offer an intuitive and flexible way to model hypothetical interventions. By isolating the marginal effects of individual features and visualizing their influence on the predicted outcome, this approach enables policymakers to prioritize resources effectively.

Finally, the integration of two models—the score-level and indicator-level frameworks—provides both interpretability and granularity. While the former highlights broad strategic areas (e.g., Housing or Jobs), the latter translates these into actionable levers (e.g., increasing room availability or secondary education attainment). This dual-layer structure enhances the interpretative power of the tool and provides practical support for evidence-based decision-making.

Chapter 4

Clustering analysis

4.1 Objective and methodological framework

The clustering analysis aimed to uncover whether distinct configurations of regional well-being can yield comparable levels of subjective life satisfaction. Specifically, we sought to determine if alternative, multidimensional pathways to well-being exist across OECD regions, each characterized by different strengths and weaknesses in key quality-of-life domains.

To achieve this, we employed the KMeans algorithm to group 411 OECD regions based on 11 standardized well-being indicators. Life satisfaction was not included in the clustering process but was later overlaid onto the results to evaluate whether similar satisfaction levels emerged from different combinations of features. A Principal Component Analysis (PCA) was used to reduce dimensionality and visualize structural differences among the clusters in two dimensions.

4.2 Clustering procedure

Data preparation and scaling

Non-numeric columns (e.g., country and region names) and the target variable *Life Satisfaction* were excluded. The remaining 11 dimensions (Education, Jobs, Income, Safety, Health, Environment, Civic Engagement, Accessibility to Services, Housing, Community) were cleaned of missing values and standardized using z-scores to ensure uniform contribution to Euclidean distance calculations during clustering.

Model implementation

We applied the KMeans algorithm with $k = 4$ clusters, selected based on interpretability and empirical coherence. The choice of the number of clusters was guided by criteria of empirical consistency, a balance between granularity and interpretability, and supported by a preliminary analysis of inertia and the elbow method. The latter revealed a visible inflection point at $k = 4$, indicating a significant reduction in within-cluster variance. Although the diagnostic plots are not reported here, this choice aligns with the principle of parsimonious clustering, whereby a limited number of semantically distinct clusters is preferable to overly fine-grained partitioning. Each cluster was then profiled based on its average scores in the input dimensions and its mean life satisfaction value.

Dimensionality reduction with PCA

PCA was applied to the scaled dataset, and the first two components were retained to facilitate a two-dimensional visualization of the cluster distribution. The resulting scatterplot revealed distinct structural groupings, highlighting both convergences and divergences in well-being profiles.

4.3 Cluster interpretation

Cluster 0: High Safety, Low Social Cohesion

Life Satisfaction Mean: 4.72
Profile: This cluster features extremely high safety scores (9.6) but suffers from low housing quality (2.8), weak civic engagement (3.3), and relatively limited community support (5.7).
Interpretation: While public safety contributes positively, the lack of social capital and inclusion limits overall subjective well-being. This group might represent regions with effective law enforcement but fragmented or underdeveloped social infrastructure.

Cluster 1: Structural and Economic Affluence, Health Fragility

Life Satisfaction Mean: 7.93
Profile: Regions in this cluster show strong performance in income (8.6), housing (8.5), and accessibility to services (8.0), but suffer from relatively poor health outcomes (4.7).
Interpretation: These regions exemplify a material and infrastructural pathway to well-being. However, the fragility in health could signal latent vulnerabilities, warranting attention in public health policy.

Cluster 2: Systemic Deprivation

Life Satisfaction Mean: 2.98
Profile: This cluster is marked by consistently low values across nearly all indicators, particularly income (0.4), housing (1.2), and jobs (3.9).
Interpretation: These are structurally disadvantaged regions with no compensatory strengths. Life satisfaction is substantially impaired, highlighting the systemic nature of deprivation.

Cluster 3: Social and Relational Well-being

Life Satisfaction Mean: 7.91
Profile: These regions score highly on community (7.9), civic engagement (7.4), safety (9.8), and health (7.7), despite modest levels of income and housing.
Interpretation: This cluster illustrates a social-relational path to well-being, where subjective satisfaction arises from strong interpersonal networks and perceived public security, rather than material wealth.

4.4 Comparative insights

A comparison of the clusters reveals several critical findings:

- **Equifinality of Well-being:** Clusters 1 and 3 exhibit nearly identical average life satisfaction despite starkly different foundations. This supports the hypothesis of multiple viable pathways to subjective well-being.
- **Systemic Vulnerability:** Cluster 2 demonstrates that the absence of strength in any domain results in significantly lower satisfaction, confirming that no single variable can compensate for multidimensional deprivation.
- **Partial Resilience:** Cluster 0, while not entirely deprived, lacks cohesive social structures and exhibits only partial well-being, despite high safety.

4.5 Visual interpretation with PCA

The two-dimensional PCA visualization reveals that clusters 1 (orange) and 3 (red), although distant in the reduced feature space, converge in terms of life satisfaction. Cluster 2 (green) is isolated, indicating substantial structural divergence from other clusters. Cluster 0 (blue) occupies an intermediate space with high internal dispersion, reflecting the heterogeneity of partial resilience.

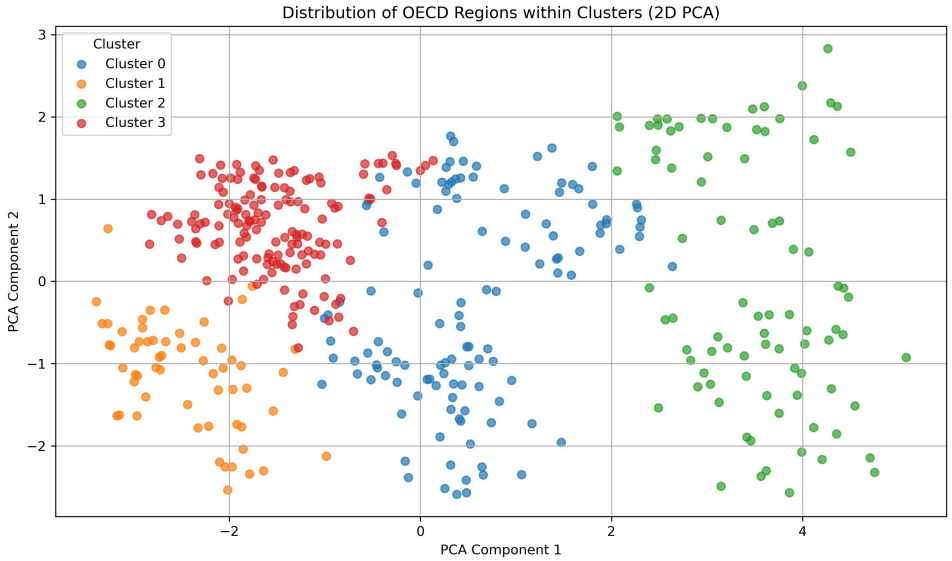


Figure 4.1: OCED Regions distribution in 2D PCA

These observations reinforce the conclusion that subjective well-being does not stem from a singular optimal configuration but can emerge from qualitatively different combinations of strengths.

4.6 Policy implications

The findings suggest that one-size-fits-all strategies for enhancing well-being are likely to be ineffective. Instead, regional policies should be tailored to local strengths and

weaknesses:

- **Cluster 0:** Initiatives to strengthen community networks and civic participation could help elevate satisfaction levels.
- **Cluster 1:** Investments in healthcare infrastructure could balance an otherwise affluent profile.
- **Cluster 2:** Multi-sectoral interventions are needed, as no single policy domain can reverse the systemic disadvantage.
- **Cluster 3:** Strategies should protect and nurture social capital while gradually improving material conditions.

These differentiated strategies underscore the necessity for a nuanced, systemic, and territorially-aware policy framework.

4.7 Conclusion

The clustering analysis reveals the existence of distinct regional well-being configurations that produce similar subjective outcomes through different pathways. It emphasizes the complex and multi-dimensional nature of life satisfaction, challenging the adequacy of aggregate scores and highlighting the value of disaggregated, profile-based policy design.

By combining clustering, PCA visualization, and interpretative profiling, this study provides a robust analytical foundation for informed and context-sensitive public policy development aimed at improving subjective well-being across heterogeneous territories.

Chapter 5

Conclusions

5.1 Answering the research question

This project set out to explore a fundamental question: *What determines life satisfaction in OECD regions, and how can we explain and predict its variation across different territorial contexts?* The analyses presented across this report provide a comprehensive, evidence-based response to this question, combining predictive modeling, interpretability techniques, and unsupervised learning in a novel and integrative framework.

First, our modeling approach established that life satisfaction is not driven by a single, dominant factor but emerges from complex, multidimensional interactions among structural indicators of well-being. Random Forest and XGBoost regressors, trained on both composite scores and raw indicators, consistently highlighted the importance of housing quality, income, community support, and employment as primary drivers of subjective well-being. Notably, these effects were not uniform but often shaped by non-linearities, threshold effects, and interaction dynamics, as revealed through SHAP analyses.

Second, by decoupling predictive inference from explanatory interpretation, the use of SHAP allowed us to precisely quantify the marginal and joint contributions of each feature to life satisfaction outcomes. These insights not only confirmed the robustness of established determinants—such as income and social capital—but also uncovered subtler effects, including the diminished or counterintuitive role of variables like education and safety in certain contexts.

In answering the research question, this project demonstrates that life satisfaction in OECD regions is determined by a complex interplay of material and relational factors, whose influence varies across territorial contexts. While economic affluence and structural infrastructure are critical in some regions, others achieve similar levels of well-being through social cohesion, safety, and civic engagement. This plurality of pathways—quantitatively validated through clustering and SHAP analysis—shows that regional life satisfaction cannot be explained through universal drivers, but must be understood as a context-specific emergent outcome shaped by trade-offs, synergies, and local strengths.

5.2 Beyond prediction: uncovering structural diversity

Beyond accurate prediction, our work emphasized the importance of structural heterogeneity across regions. The clustering analysis demonstrated that similar levels of life satisfaction can be achieved through fundamentally different well-being configurations—a phenomenon known as *equifinality*. Two distinct clusters, one rooted in economic affluence and another in social cohesion, exhibited comparable life satisfaction scores, yet were underpinned by divergent strengths and vulnerabilities.

This finding is crucial: it challenges the assumption of a single "optimal" path to well-being and underscores the necessity of place-based, multidimensional strategies. Moreover, the identification of a severely deprived cluster—characterized by structural deficits across all dimensions—highlights the moral and policy urgency of addressing systemic disadvantage.

5.3 Implications for research and policy

From a scientific perspective, this project demonstrates the power of integrating machine learning with explainable AI and unsupervised learning to address complex social science questions. The methodology applied here goes beyond descriptive statistics or composite indices, offering a transparent and empirically grounded lens to explore latent patterns in human well-being.

From a policy standpoint, the results advocate for a radical shift in how life satisfaction is approached at the regional level. Rather than promoting uniform prescriptions based on universal rankings, the insights generated here call for tailored interventions that reflect local configurations of strength and need. The dual use of aggregate scores and granular indicators—each modeled and interpreted separately—enables a flexible toolkit for both strategic prioritization and operational planning.

5.4 Final remarks

In conclusion, this study reaffirms the multidimensional and contextual nature of well-being. It shows that subjective life satisfaction cannot be reduced to any single variable or ranking, but must be understood as the emergent property of multiple interlocking domains. Through robust modeling, transparent interpretation, and structural clustering, we provide both theoretical insight and practical guidance for understanding and improving life satisfaction across OECD regions.

Ultimately, this project demonstrates that data science is not merely a technical tool but a transformative approach to uncovering the hidden structures that shape human flourishing. It invites scholars and policymakers alike to embrace complexity, resist oversimplification, and commit to context-sensitive, evidence-based pathways toward a better quality of life for all.