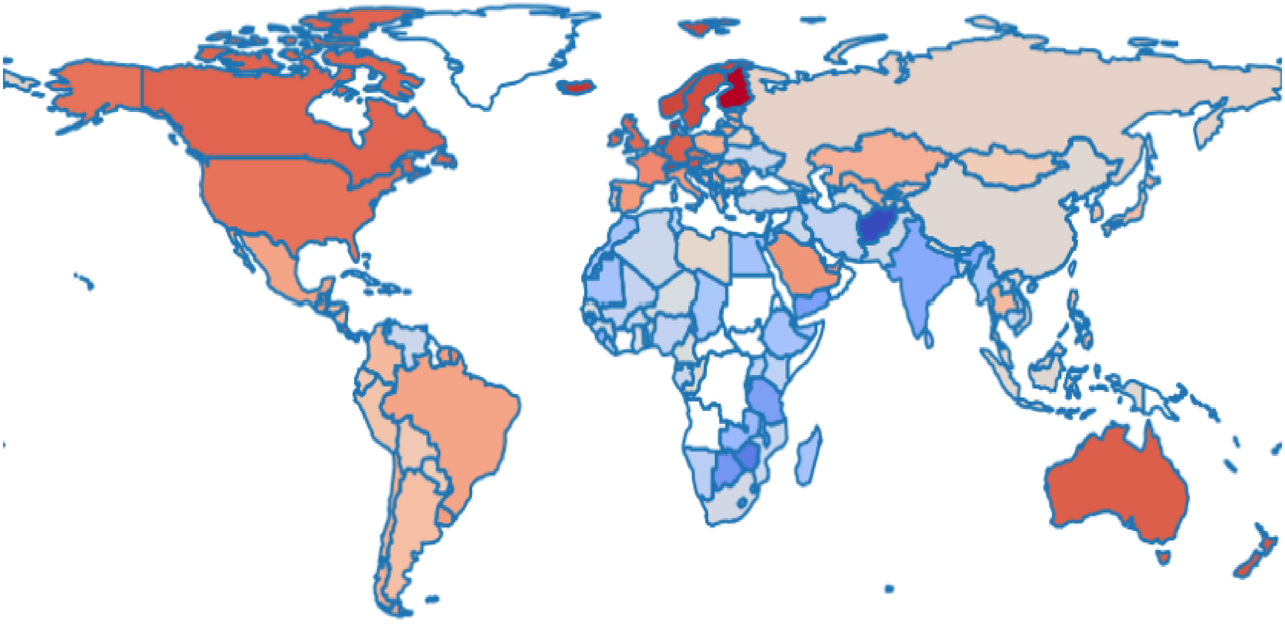
DataScientest

Data Analytics Project:
World Happiness



by

Evangelos Ziogas
Delphine Parmentier

Project Mentor:
Tarik Anouar

August 2024

Table of content

Introduction

In recent years, the importance of happiness and well-being has gained significant attention, particularly since the COVID-19 pandemic. This global crisis has highlighted the critical need for mental health support, social connection, and overall life satisfaction, since people worldwide faced unprecedented challenges and disruptions. This increased awareness has led to prioritizing and placing greater emphasis on these aspects of life.
As a result, more and more research projects dedicated to understanding and improving well-being are conducted globally. Among these, the World Happiness Report stands out as a significant annual publication that assesses the state of global happiness by evaluating factors such as income, social support, life expectancy, freedom, generosity, and corruption.

The following project will give a statistical insight into the World Happiness index, a study conducted in the above-mentioned *World Happiness Report,* arising from a partnership of Gallup, the Oxford Wellbeing Research Centre, the UN Sustainable Development Solutions Network, and the WHR's Editorial Board.

Objectives

Our project on world happiness aims to explore and analyze the various factors that contribute to the well-being and satisfaction of people across different countries. By examining data on economic, social, and environmental indicators, as well as individual perceptions of happiness, we seek to uncover the underlying drivers of happiness and understand how these elements interact.

Our analysis will therefore measure to what extent these factors contribute to a nation's happiness, their relative importance in the citizen's well-being, and will also draw comparisons between the individual countries.

Methodology for Data Pre-Processing

1. Data Collection

Individual's self assessment of their lives were collected through surveys distributed in more than 150 countries in the world. These data are split in two datasets[1]: one showing Happiness Rates for the year 2021, and the other Happiness Rates from 2005 to 2020.

The same data were also reported in the 2021 dataset for a hypothetical country named "Dystopia". This country, showing the world's lowest national averages among all measurements, will be considered the least happy nation and will serve as a point of comparison to the others.

2. Data Exploration

Our analysis of happiness rates across the world is based on several important measurements, mainly represented by the below 6 variables:

Logged GDP per capita	Normalized measure of the economic production of a country
Social support	Refers to assistance or support provided by members of a social network to an individual (ie, social programs/support from family, etc.)
Healthy life expectancy	Average lifespan of citizens in good health, without any incapacity or limitation in daily life
Freedom to make life choices	Ability of individuals to make decisions about their own lives without undue restrictions or limitations
Generosity	Level of charity in a nation (ie, amount of donations performed)
Perceptions of corruption	Extent to which people think corruption exists within their governments and institutions

It is worth noting that these factors do not influence the overall happiness score reported for each country, but they do clarify why some countries rank higher than others. Therefore, those are considered **explanatory variables** of our dataset.

Our **target variable** is represented by the Ladder Score. This score is a subjective measure resulting from individuals rating their overall well-being through the Cantril Self-Anchoring Scale[2], a tool used extensively by Sociology Researchers worldwide. This scale asks respondents to rate their current life situation on a ladder where the steps represent the best and worst possible lives they can imagine. As such, this self-evaluation also includes a personal assessment of both positive and negative emotions, reflecting a comprehensive view of their emotional and psychological state.

The following additional variables were also included in our datasets initially, but we decided to remove them for sake of relevance and clarity while manipulating the data:

- Standard error of ladder score
- Upper and lower whiskers
- Positive and negative affect

Since standard error and plots' whiskers can be generated at any time during analysis, we decided to remove the corresponding columns and rather generate those data if required. Positive and negative emotions being intrinsic measures of the ladder score, the ladder score alone is therefore sufficient to be considered for this analysis, as it reflects both parameters.

3. Data Cleaning

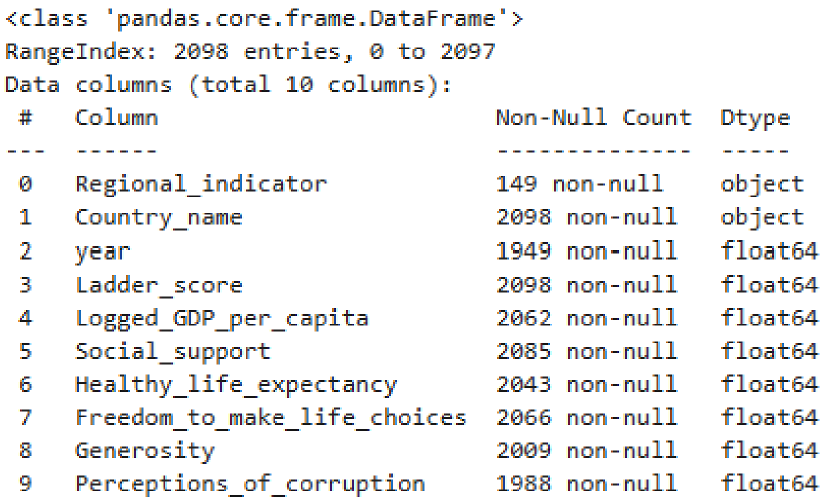
After properly identifying the variables of importance for the analysis, we proceeded with merging the data from the years 2005-2020 with that of 2021 to create a single and complete dataset, and then started the data cleaning process:

- **Variables removed:**
  - Data related to Dystopia were grouped into a separate dataset. This country now has its own data bank, should it be needed for comparison.
  - Standard error of ladder score, upper and lower whiskers, positive and negative affect were removed from the dataset (see ¶ *Data Exploration*).

- **Data normalization:** while most of the data fell within the 0-1 range, it appeared that logged GDP per capita, healthy life expectancy, and generosity had values outside this range and therefore required normalization. The Min-Max method was used to scale these variables.

- **Duplicates:** none were identified in our merged dataset.

- **Missing values:** the expected number of entries in the merged dataset being 2098, 8 out of 9 variables in total were identified as being incomplete (fig.1):



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2098 entries, 0 to 2097
Data columns (total 10 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   Regional_indicator            149 non-null    object 
 1   Country_name                  2098 non-null   object 
 2   year                          1949 non-null   float64
 3   Ladder_score                  2098 non-null   float64
 4   Logged_GDP_per_capita         2062 non-null   float64
 5   Social_support                2085 non-null   float64
 6   Healthy_life_expectancy       2043 non-null   float64
 7   Freedom_to_make_life_choices  2066 non-null   float64
 8   Generosity                    2009 non-null   float64
 9   Perceptions_of_corruption     1988 non-null   float64
```

***Fig.1:** Overview of the initial merged dataset after removing unnecessary columns.*

Most of the variables of this dataset are quantitative, except regional indicators and country names, which are categorical. Here is how we have addressed the missing values for each affected category:

- Regional indicator: this column has been populated with the assistance of AI. Initially, we completed the missing values based on the existing country-region pairs of the dataset. After this step, any remaining country without a corresponding region was mapped to the most appropriate region available in the dataset, thanks to the AI.

- Year: no missing values were identified in the initial 2005-2020 dataset for the year, meaning that any missing years must be from 2021. This column was populated based on this logic.

- Other remaining quantitative variables: to determine the best method for replacing missing values - either by the mean or the median - we assessed the distribution of each variable (fig2). Given the asymmetrical distributions observed, we decided to impute the missing values using the median for each variable. This approach is more robust to skewed distributions and less sensitive to outliers and extreme values. Each column was populated according to this logic, resulting in a final clean and complete merged dataset (fig.3).

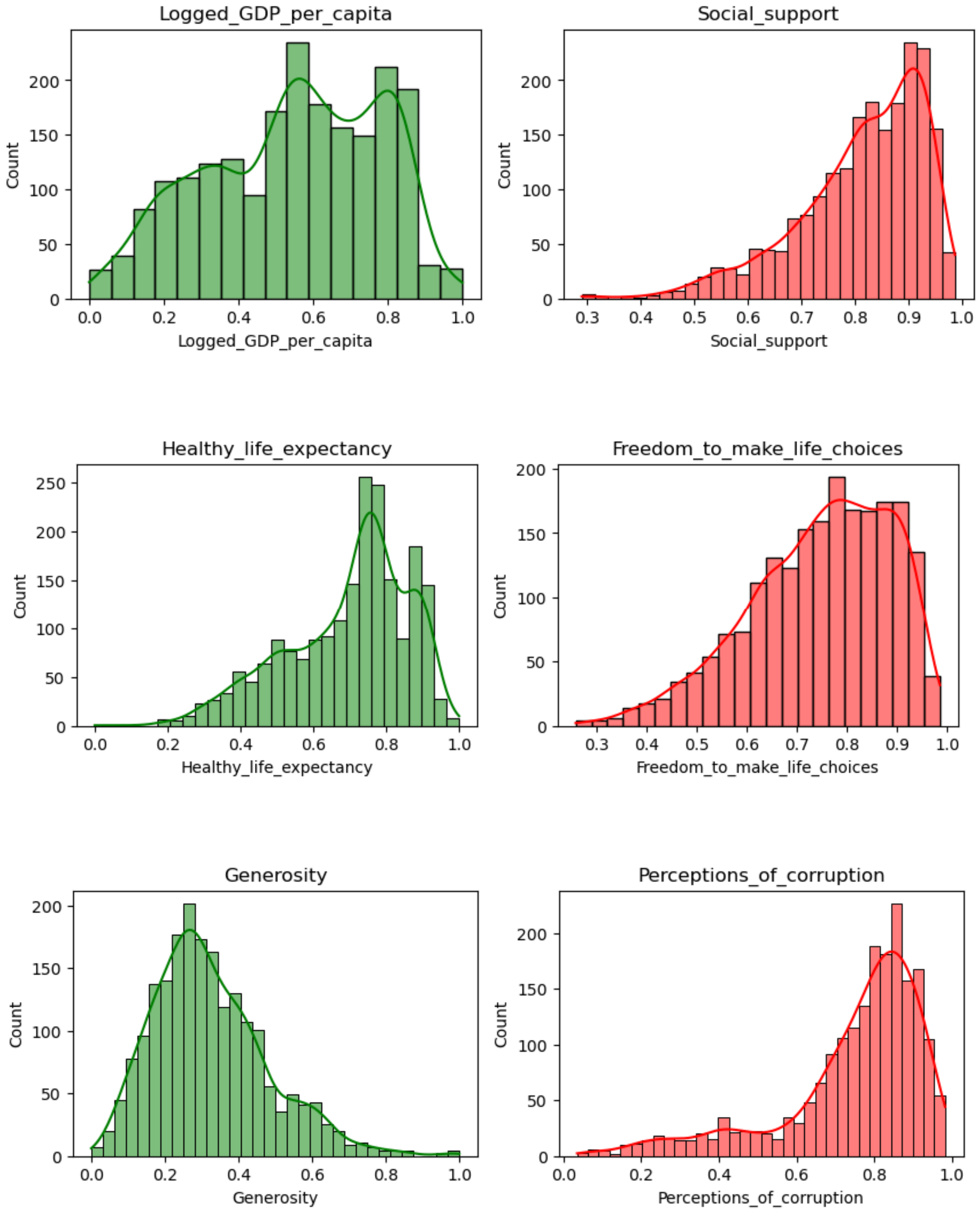
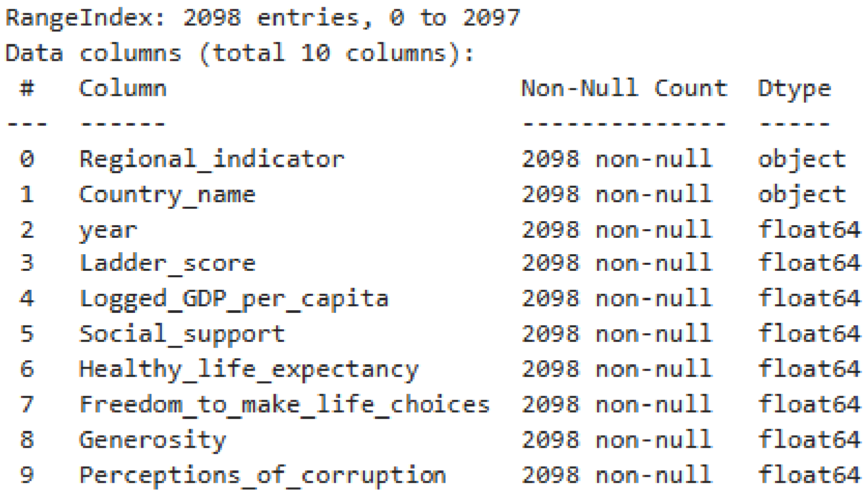


Fig.2 : *Distribution of the quantitative variables affected by missing values (logged GDP per capita, social support, healthy life expectancy, freedom to make choices, generosity and perception of corruption).*



```
RangeIndex: 2098 entries, 0 to 2097
Data columns (total 10 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   Regional_indicator            2098 non-null   object 
 1   Country_name                  2098 non-null   object 
 2   year                          2098 non-null   float64
 3   Ladder_score                  2098 non-null   float64
 4   Logged_GDP_per_capita         2098 non-null   float64
 5   Social_support                2098 non-null   float64
 6   Healthy_life_expectancy       2098 non-null   float64
 7   Freedom_to_make_life_choices  2098 non-null   float64
 8   Generosity                    2098 non-null   float64
 9   Perceptions_of_corruption     2098 non-null   float64
```

Fig.3: *Overview of the final merged dataset after completing pre-processing steps.*

Data Visualization

1. Analysis of the happiness score across countries

First, we examined the distribution of the ladder score across all countries from 2005 to 2021 (fig. 4).

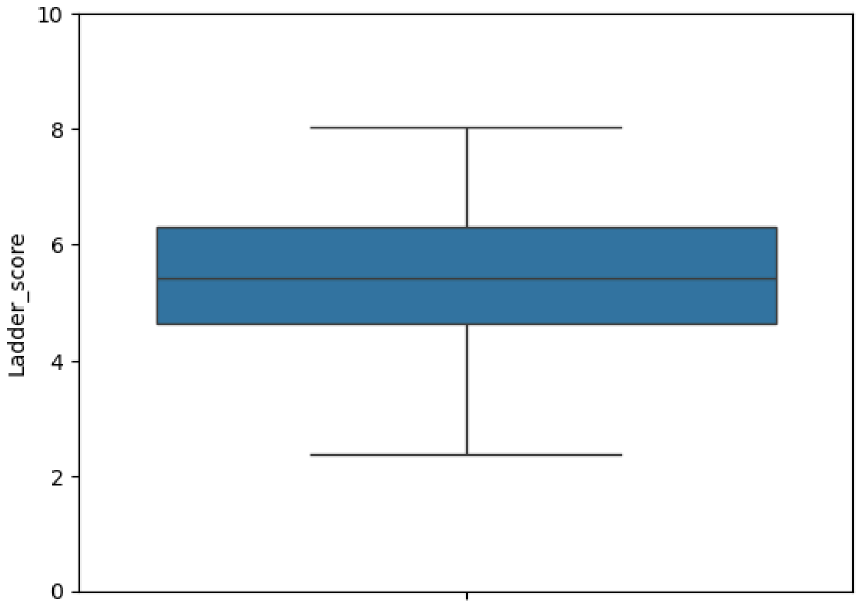


Fig.4. *Distribution of the ladder score values across all countries (2005-2021).*

The worldwide distribution of the ladder score appears symmetric, indicating a balanced dataset that fairly represents the well-being variable, encompassing a diverse range of individuals with varying levels of reported happiness.

Then, we analyzed rankings across continents (fig.5) and ultimately identified the top 10 countries ranked as the happiest in the world in 2021 (fig.6).

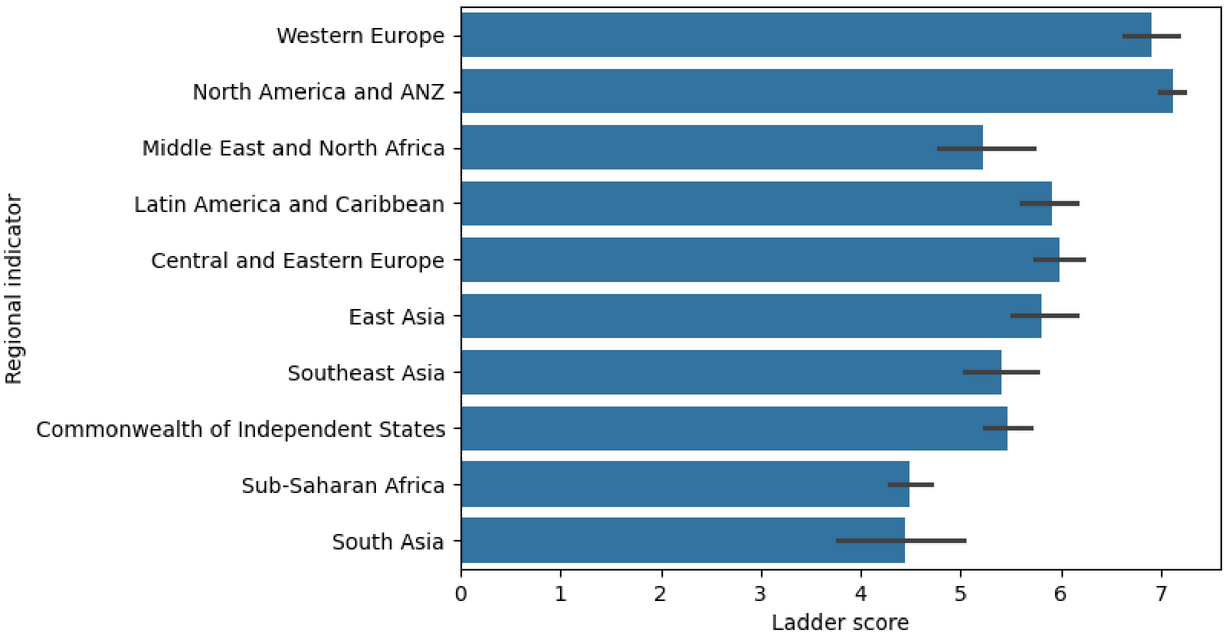


Fig.5: *Ranking of the happiest continents in 2021.*

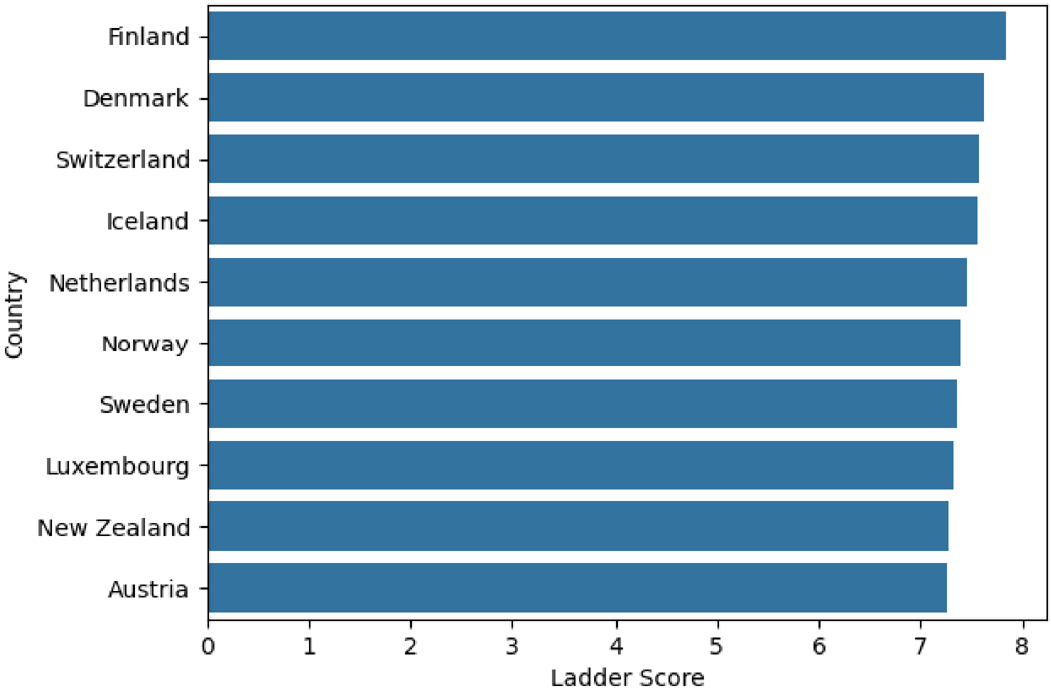


Fig.6: *Top 10 happiest countries in the world in 2021.*

As we can see, Western Europe, North America and New Zealand are ranked among the happiest regions in 2021, with Finland and Denmark being the leading countries.

To gain a comprehensive global perspective on the evolution of the scores, we also examined the historical trend from 2005 to 2021 (fig. 7).

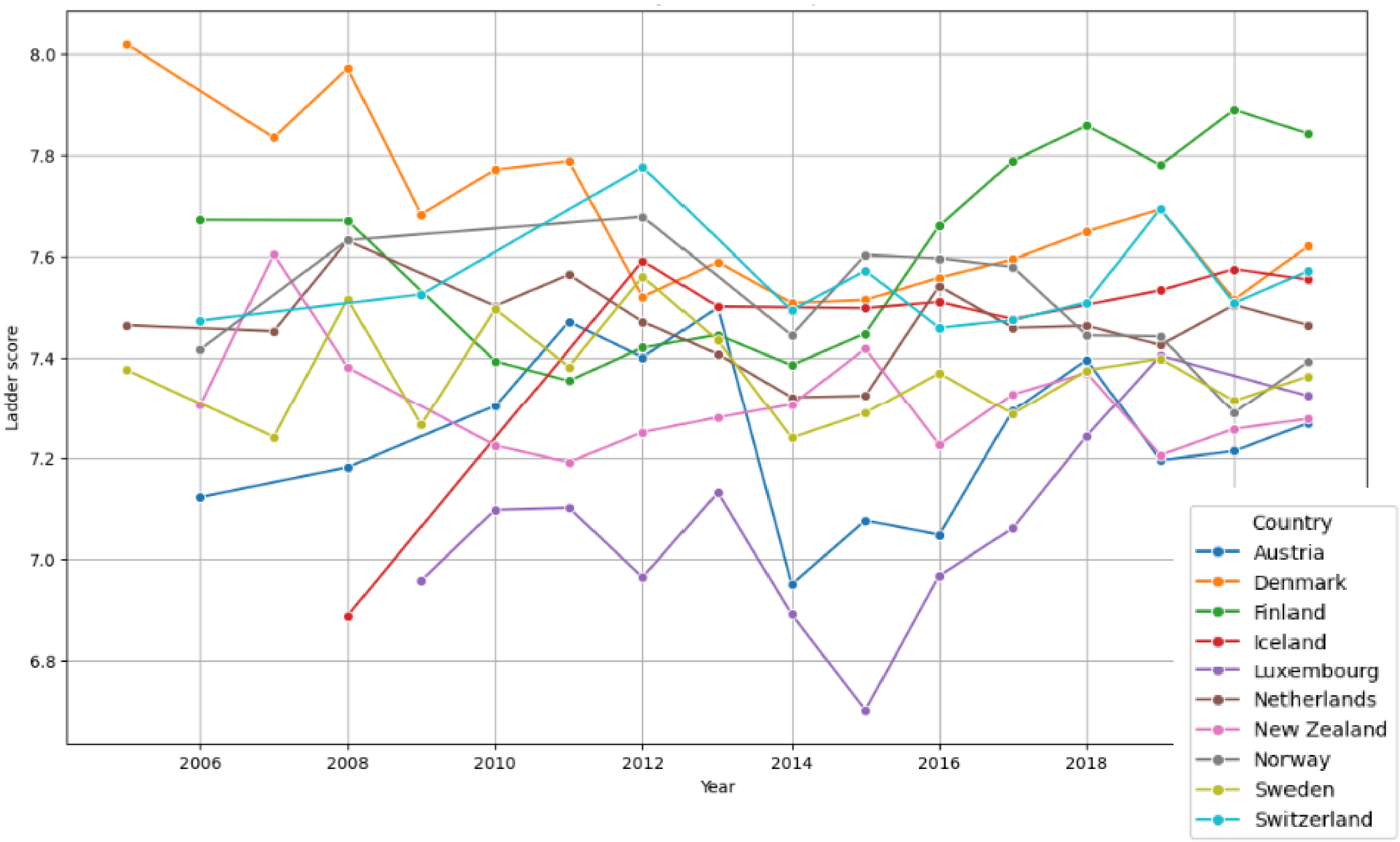


Fig.7: *Ladder score values over years for the top 10 countries in 2021.*

While Denmark and Finland still appear to be among the happiest countries over time, we can also observe from this timeframe that a global decrease of the ladder score was observed for all countries in 2014.
One might attribute this decline to socio-economic factors, leading to instability, economic hardship, and insecurity worldwide.

We have also compared the 10 least happy countries with the hypothetical nation of Dystopia (fig.8). Dystopias' average Ladder Score is set as 2.43. It is represented in the following barplot as a red line. Nations like Afghanistan, Zimbabwe and Rwanda are extremely close to the Ladder Score of Dystopia. Most of the countries depicted here suffer from political turmoil and severe economic conditions, meaning they are very close to the hypothetical worst nation.

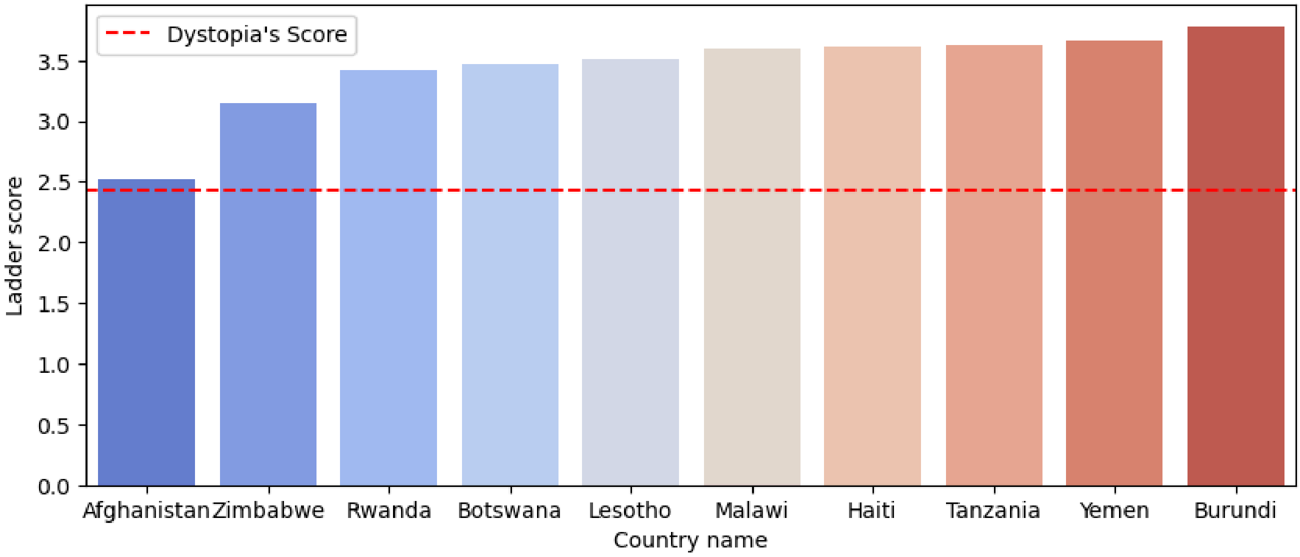


Fig.8: *Comparison of the 10 least happy countries with Dystopia.*

2. Relationship of the happiness score to external factors

As a next step, we decided to focus on understanding the specific factors that might influence happiness rates. To achieve this, we used a heatmap to analyze the correlations between the ladder score and each of the six factors included in this study. It appears that the two predominant factors are logged GDP per capita and healthy life expectancy (fig.9).

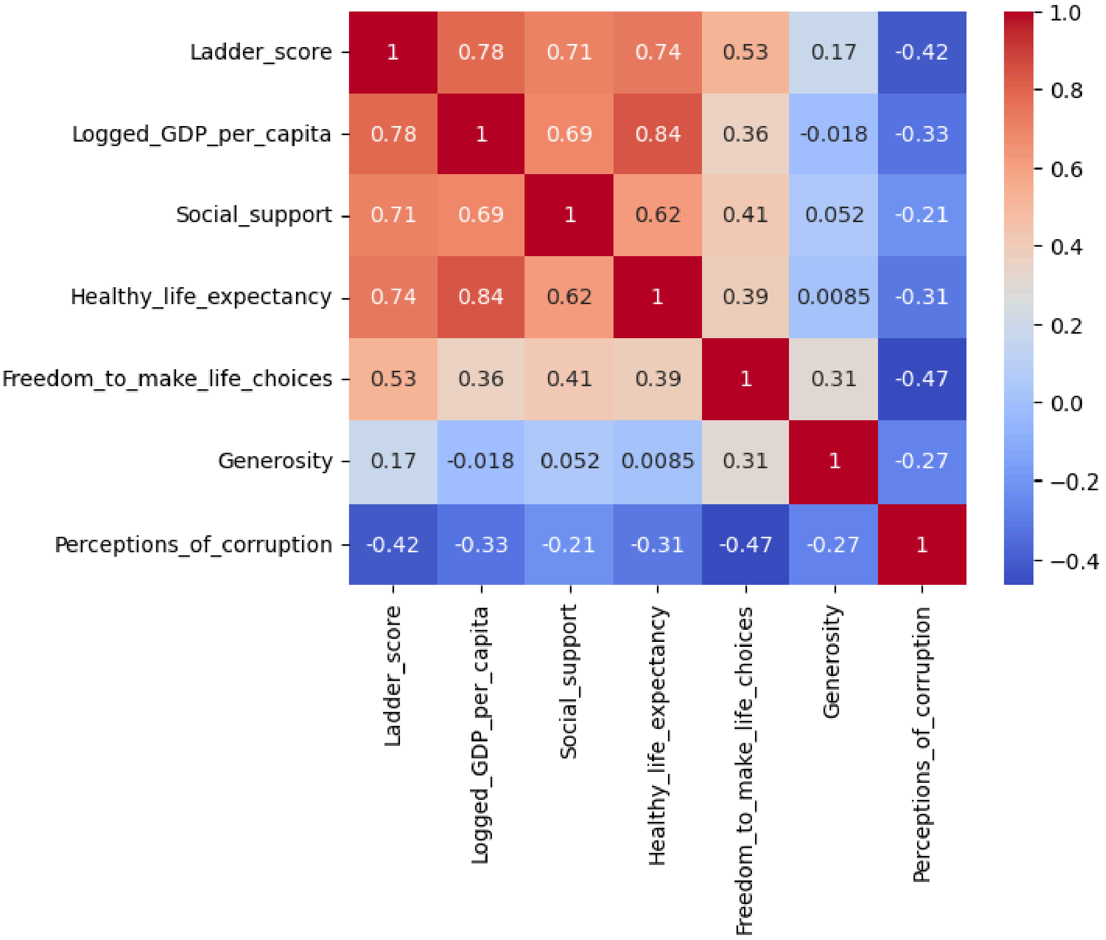


Fig.9: *Correlation between the 6 factors known to influence life evaluation (2005 - 2021).*

In order to examine the relationship between the ladder score and these two identified factors, we applied a Pearson statistical test. Our goal was twofold: first, to determine whether logged GDP per capita and life expectancy significantly impact the ladder score, and second, to assess the linearity of the relationship between these factors and our target variable.
In both cases, the resulting p-value was below the standard threshold of 0.05 (0 in each case), suggesting a strong and highly significant correlation of each factor with the ladder score.

The calculation of the correlation coefficient confirmed this assumption, yielding values of 0.74 for healthy life expectancy and 0.78 for logged GDP per capita, indicating a strong linear relationship of each factor with the ladder score.

Upon comparing the healthy life expectancy across the three happiest countries and the three least happy countries, a striking disparity emerges. The happiest nations exhibit life expectancies that surpass those of the unhappiest counterparts by a significant margin, typically ranging from 10 to 20 years (fig.10). This discrepancy suggests substantial differences in healthcare accessibility, overall quality of life, and potentially socio-economic conditions between these distinct groups of countries.

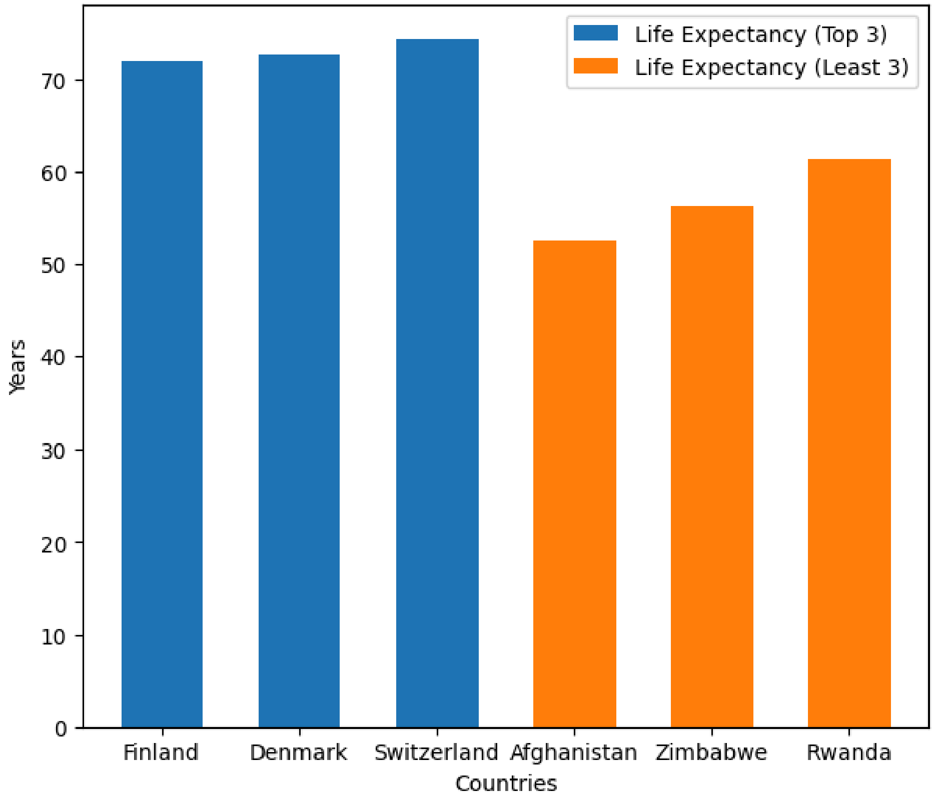


Fig.10: *Comparison of Life Expectancy between the most and least happy countries.*

Our analysis of the logged GDP per capita across the three happiest countries and the three least happy countries reveals a notable disparity as well, of approximately 8 to 10 points (fig. 11). This discrepancy underscores significant economic inequality between these nations.

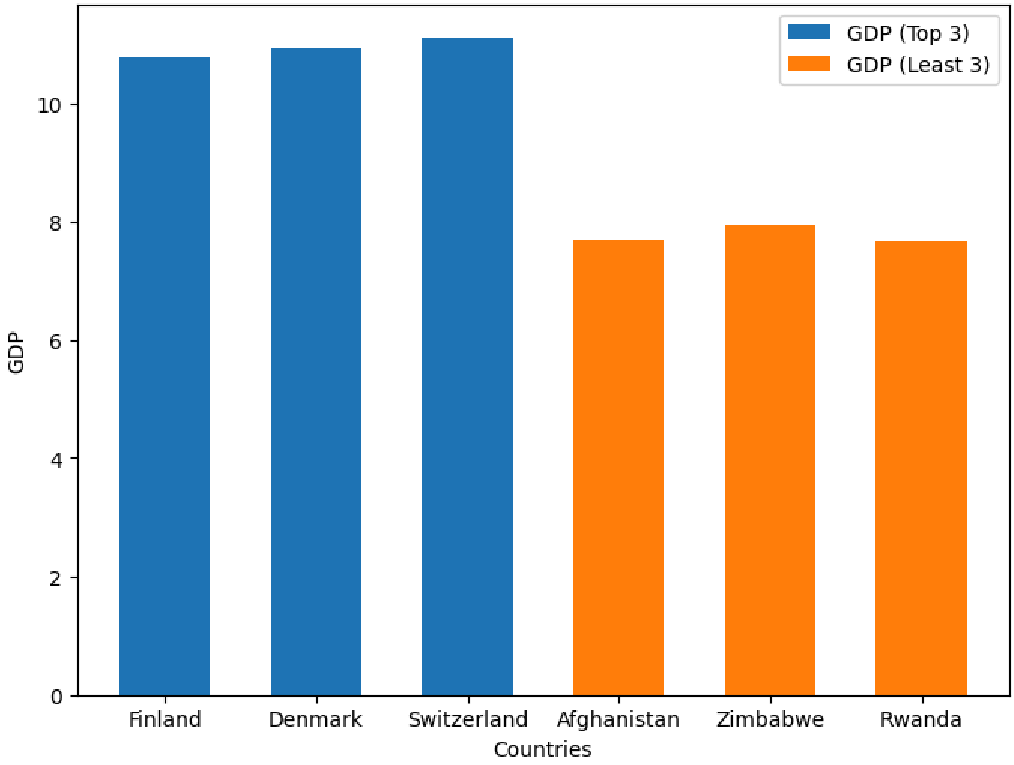


Fig.11: *Comparison of economic rates between the most and least happy countries.*

Data Modelling

This section is dedicated to predicting global happiness levels, thanks to different machine learning models. The high predictive capability of machine learning models offers valuable insights for researchers and organizations to better understand human well-being. By revealing patterns and trends across different countries, these models contribute to enhancing the quality of life and addressing global disparities.

In this project we applied the four following models: linear regression, decision tree regression, random forest and gradient boosting. We assessed the performance of each model in delivering reliable forecasts by examining their respective metrics.

These models were applied to the preprocessed data (see ❡ *Methodology for Data Pre-Processing*), with an additional encoding step to ensure suitability for the model. After data encoding, the following main steps were systematically implemented:

- Dividing the dataset into two parts: one dedicated to training (80% of the data), one dedicated to evaluation (20% of the data). This is to ensure the model is able to perform not only on familiar data but also on new data, thus preventing overfitting.

- Training the model on the train dataset and assessing its accuracy by analyzing the coefficient of determination, R².
- Predicting on both train and test data, and assessing the model's performance by analyzing the related metrics: MAE, MSE, RMSE.

  - **MAE - Mean Absolute Error:** is the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.
  - **MSE - Mean Squared Error:** is the average of the squared difference between the actual and predicted values in the data set. It is an overall measure of the variance of the residuals.
  - **RMSE - Root Mean Squared Error:** is the square root of the mean square error. It measures the standard deviation of the residuals.

The outcome of the predictions and associated performance results are presented for each model in the following paragraphs.

1. Linear Regression

a. Training

Linear models have the advantage of being among the simplest machine learning models. The purpose here is to represent the link between explanatory and target variables, by fitting a linear equation to the data that best predicts the target variable based on the explanatory variables.

The following coefficient of determination R² have been obtained after training the model:

- R² value of the train set : 0.75
- R² value of the test set: 0.73

With the train set capturing 75% of the ladder score variance and the test set capturing 73%, we can conclude that the model is relatively effective at predicting the ladder score based on the explanatory variables. The remaining 25-27% of variance are not explained by the model and probably due to randomness.
Additionally, both R² values are quite similar, suggesting that the model is not overfitting and performing consistently across both datasets.

b. Prediction

After proceeding with prediction on both the train and the test sets, we examined the related performance metrics:

	Train dataset	Test dataset
MAE	0.43	0.45
MSE	0.31	0.33
RMSE	0.56	0.58

The low values observed for the three types of metrics indicates that the predictions are reasonably accurate, with a relatively low level of errors. Additionally, those values appear to be quite close from each other, suggesting that the model generalizes well from the train data to the test data.

To visually assess the accuracy of the model's predictions against the actual outcomes, a scatter plot of the predicted values versus the true values has been generated (fig. 12). Overall, we can see the true values are relatively close to the regression line, meaning that the prediction is good.

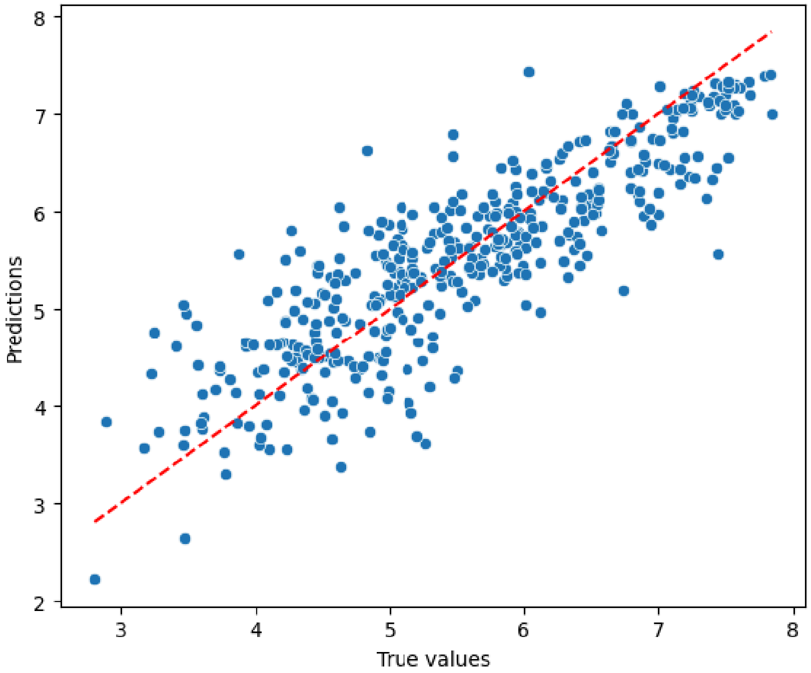


Fig.12: *Comparison of actual values vs predicted values.*

This was further validated by plotting the residuals to analyze the distribution of errors (fig. 13 a-c).

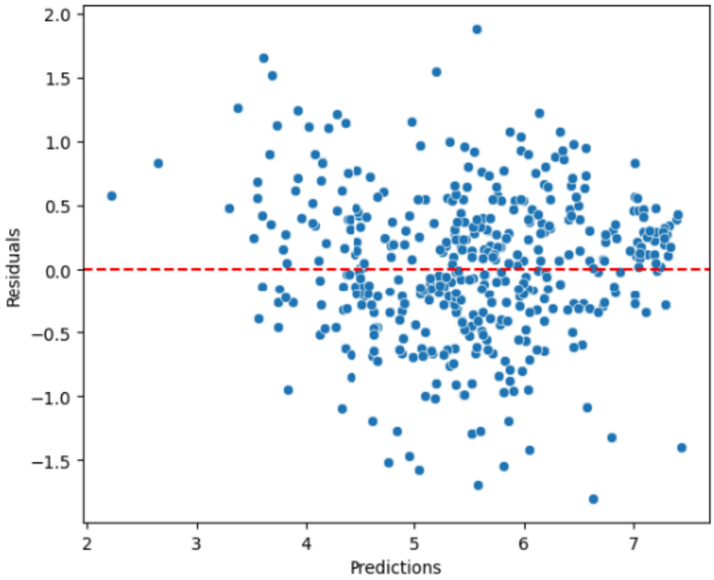


Fig. 13-a: *Scatter plot of residuals.*

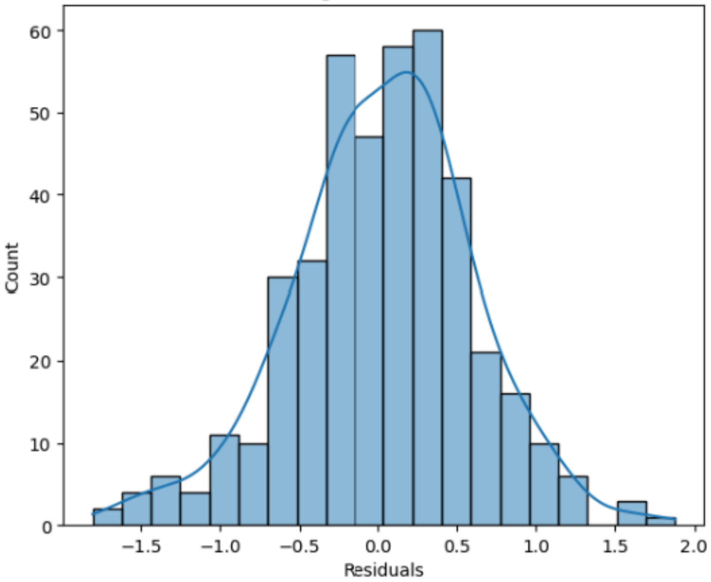


Fig. 13-b: *Histogram of residuals.*

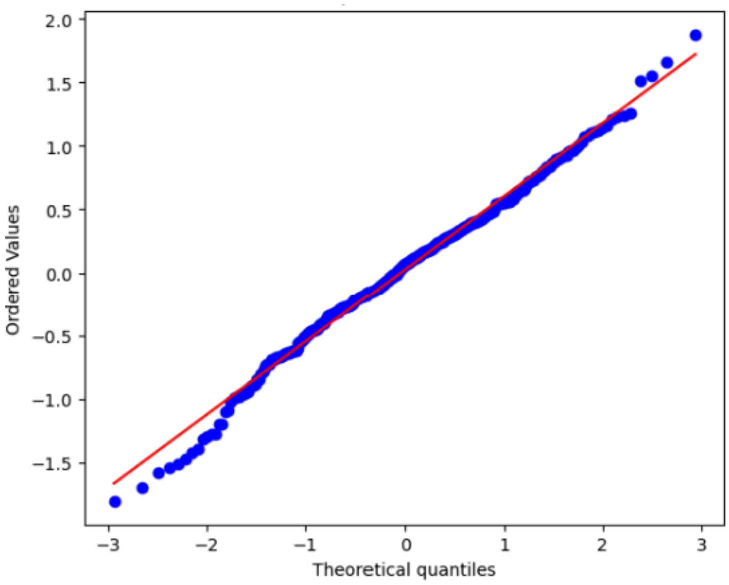


Fig. 13-c: *QQ plot of residuals.*

- The scatter plot (fig. 13-a) shows that the residuals are randomly distributed around the horizontal axis (zero line), meaning that the errors are randomly distributed and that the model is appropriately capturing the relationship between variables.
- The histogram (fig. 13-b) shows a normal distribution of the residuals, suggesting that the model is performing well with few or no outliers present in the datasets.
- Finally we can see on the QQ plot (fig. 13-c) that residuals are lying close to the 45-degree line, consistent with the previous normal distribution observed on the histogram.

Overall, we can conclude this model is quite performant.

2. Decision Tree

a. Training

Decision trees are simple models that create branching structures based on decisions related to the target variable. Decision trees are great when it comes to capturing relationships between variables, even when they are non linear. Their visual representation resembles a tree, making it easy to understand how decisions unfold. The hyperparameters chosen for the decision tree were max_depth = 3 and min_samples_leaf = 25. By limiting the depth of the tree to 3, the risk of overfitting the train dataset is reduced. The leaf sample size of 25 was chosen to help combat overfitting. Additionally the number 42 was chosen for the randomness so that the results are consistent.

The following coefficient of determination R² have been obtained after training the model:

- R² value of the train set : 0.74
- R² value of the test set: 0.69

The scores are quite favorable, exceeding 70% and closely clustered together, which suggests that the model is not overfitting. However the difference between the R² value in the train set and the R² value in the test set is quite high, meaning that this model is probably not suitable for our data.
In such cases, we aim for better consistency between training and test performance to ensure robust generalization.

The features importance of our decision tree is consistent with the previous heatmap (see ¶ *Relationship of the happiness score to external factors)*. Here too, the logged GDP per capita and healthy life expectancy emerge as the two factors with the greatest influence on the model (fig. 14). For the decision tree the regional indicator plays little role.

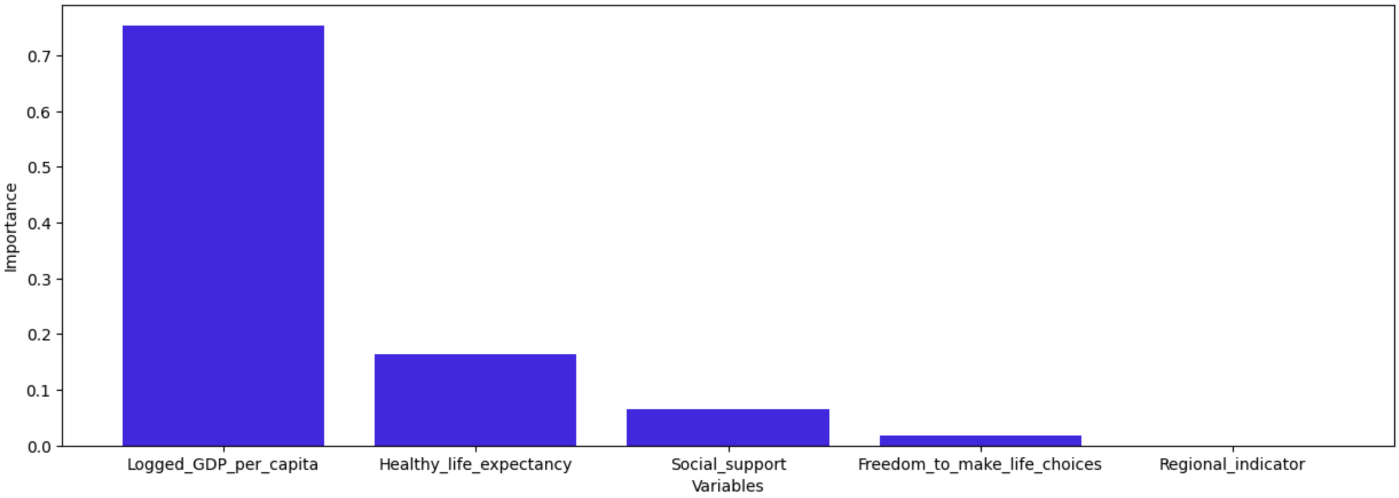


Fig.14: *Top 5 features of importance (Decision Tree).*

b. Prediction

For the prediction of the train set and the test set of the Decision tree we have concluded following:

	Train dataset	Test dataset
MAE	0.44	0.47
MSE	0.31	0.37
RMSE	0.56	0.60

The above metrics for the decision tree also have reasonably low values, showing that this model is quite accurate.

When it comes to the comparison between the predicted and the actual values, we observe gaps in the scatterplot (fig.15). This implies that the model predictions often fail to capture the true values.

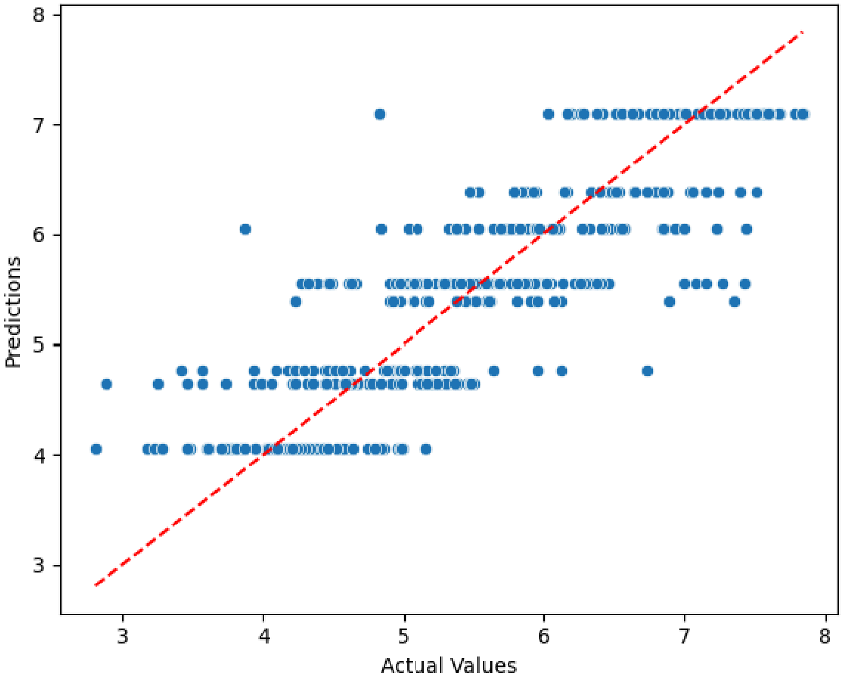


Fig.15: *Comparison of actual values vs predicted values.*

Consistently with this observation, analyzing the distribution of errors enabled to highlight the following:

- The dispersion of the residuals (fig. 16-a) is showing gaps in its distribution, highlighting the same problem depicted in (fig. 15).
- The histogram (fig. 16-b) appears to be slightly deviating from the standard bell-shaped curve, with asymmetrical left and right tails and left skewness, suggesting more frequent occurrences of negative residuals than expected in normal distribution.
- The QQ-plot (fig. 16-c) is showing some deviations as well, with residuals appearing below the bottom and above the top of the reference line. This indicates lighter tails in the lower quantiles and heavier tails in higher quantiles compared to theoretical distribution. This is aligned with previous deviation from normal distribution (fig 16-b) and suggests that the model is not adequately capturing the data variability in lower and upper ends.

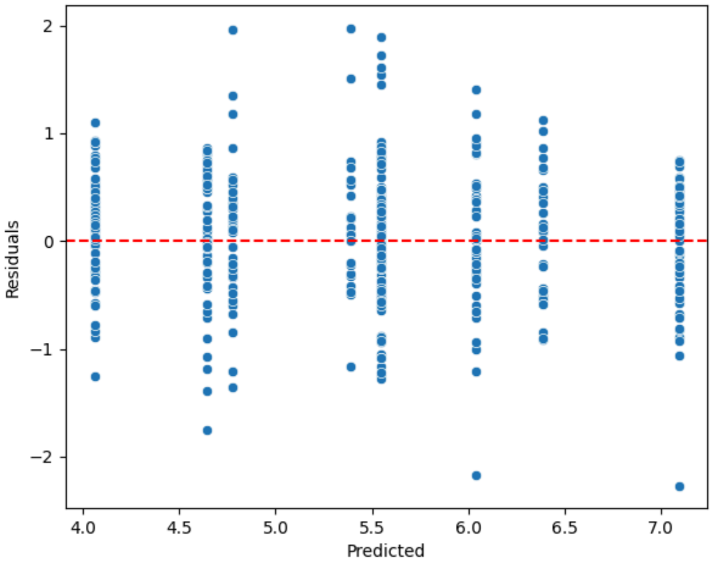


Fig. 16-a: *Scatter plot of residuals.*

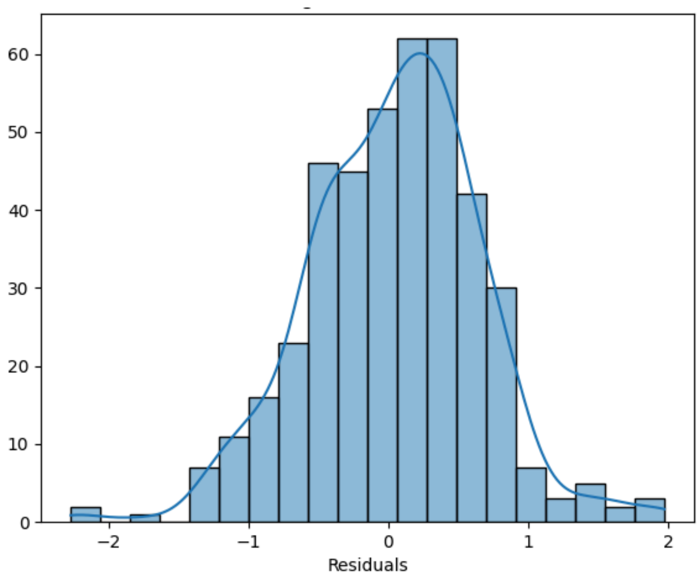


Fig. 16-b: *Histogram of residuals.*

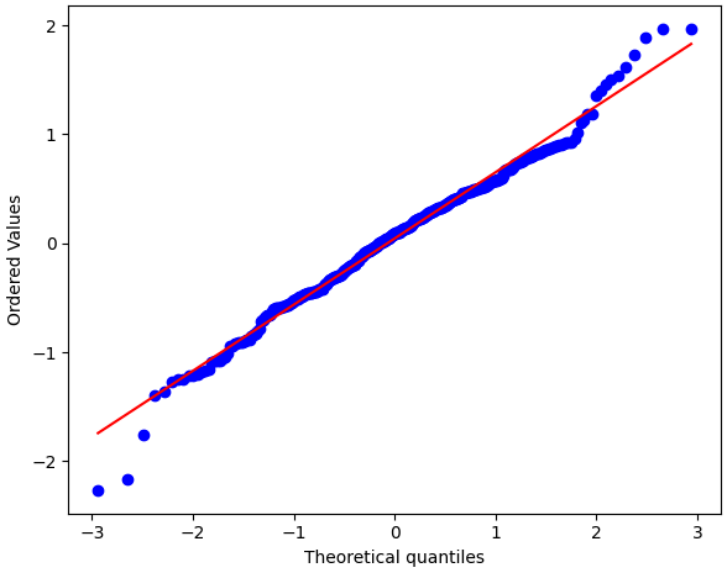


Fig. 16-c: *QQ plot of residuals.*

3. Random Forest

a. Training

Random Forest constructs decision trees in parallel, with each tree being trained independently of the others. The final prediction of this model is made by averaging the predictions of all trees, hence reducing variance and preventing overfitting.

While a single decision tree offers a straightforward approach to modeling, Random Forest introduces additional complexity by aggregating multiple decision trees to enhance performance and robustness. Similar to the Decision Tree, the hyperparameters chosen for the Random Forest were max_depth = 3 and min_samples_leaf = 25. These settings help avoid overfitting. The number 42 was also chosen for the randomness, so that the results are consistent.

The following coefficient of determination R² have been obtained after training the model:

- R² value of the train set : 0.77
- R² value of the test set: 0.73

The scores here also are quite favorable, exceeding 70 and closely clustered together, which suggests that the model is not overfitting. The Random forest model exceeds the scores of the decision tree and the linear regression.

The score values are also very close to each other which indicates that there is no overfitting.

The Random forest again confirms the feature importance established in the above heatmap and previous models (fig.17). The logged GDP per capita and healthy life expectancy are consistently the two variables with the greatest influence on the model.

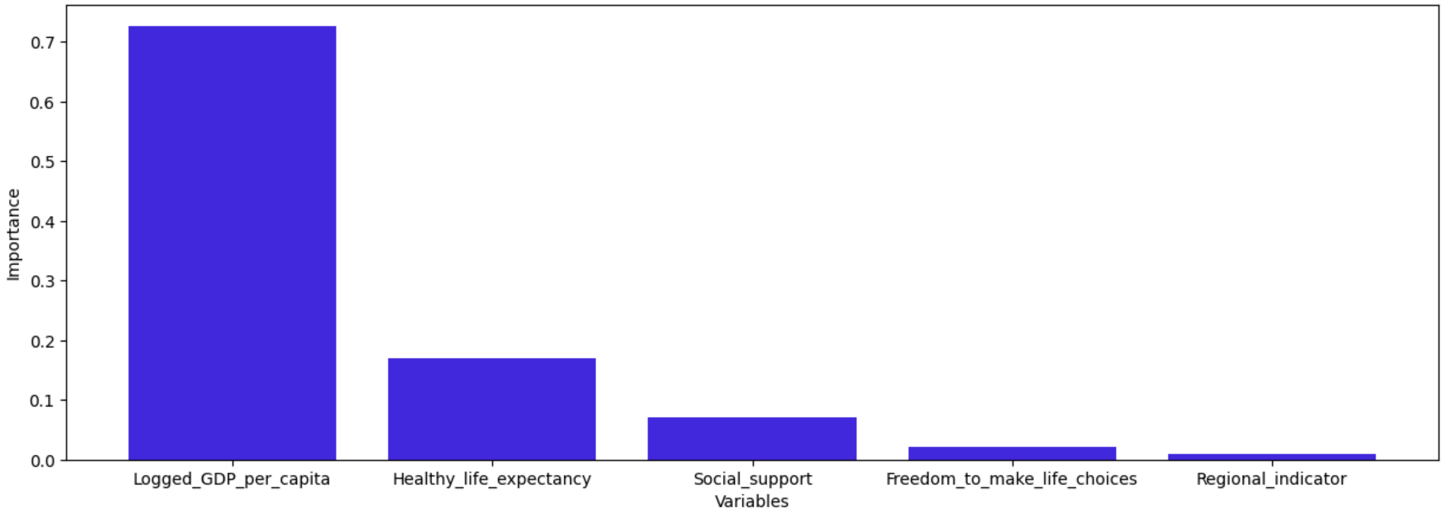


Fig.17: *Top 5 features of importance (Random Forest).*

b. Prediction

Predictions for the Random Forest model led to the following metrics and conclusion:

	Train dataset	**Test dataset**
MAE	0.41	0.45
MSE	0.28	0.33
RMSE	0.53	0.57

This model shows slightly better values than the decision tree and the linear regression models.

On the scatterplot of the actual vs predicted values (fig.18), we can observe a pattern of the true values with slight gaps around the red line (predicted values).

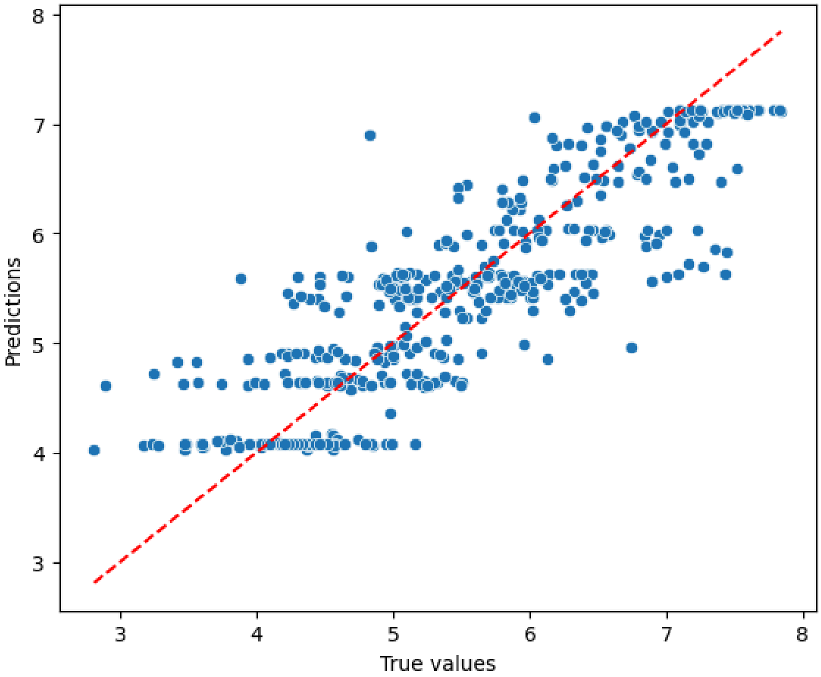


Fig.18: *Comparison of actual values vs predicted values.*

.

The same kind of pattern appears when plotting the residuals (fig. 19-a). Additionally, the normal distribution on the histogram is slightly deviating to the left (fig. 19-b), indicating inaccuracies in our model. The QQ plot is in agreement with previous observations, with residuals mostly aligning on the 45-degree line (fig. 19-c), with the exception of some extreme values behind the reference line in the lower quantile and above the reference line in the upper one.

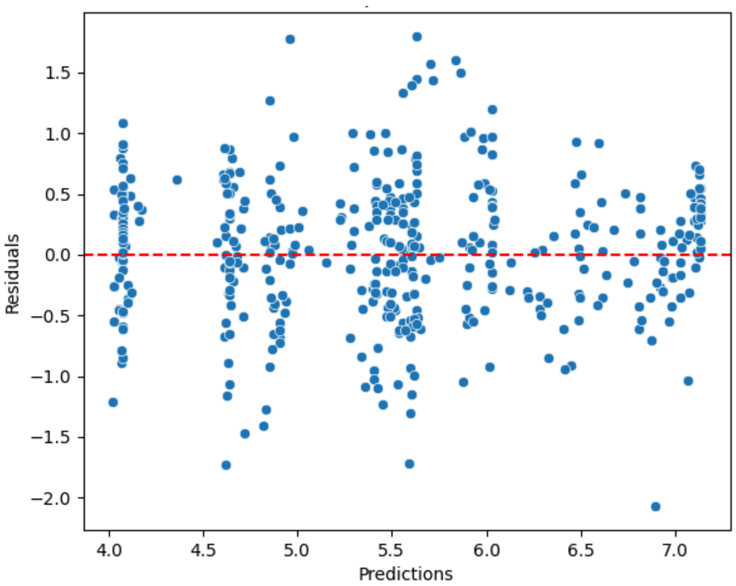


Fig. 19-a: *Scatter plot of residuals*

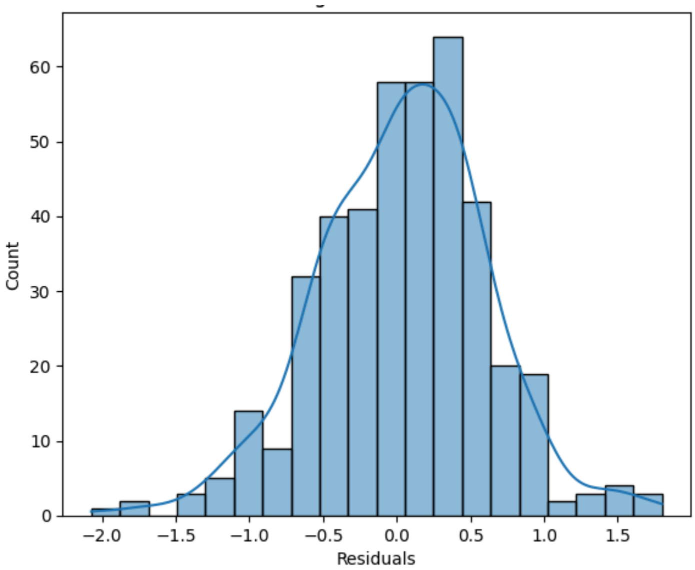


Fig. 19-b: *Histogram of residuals*

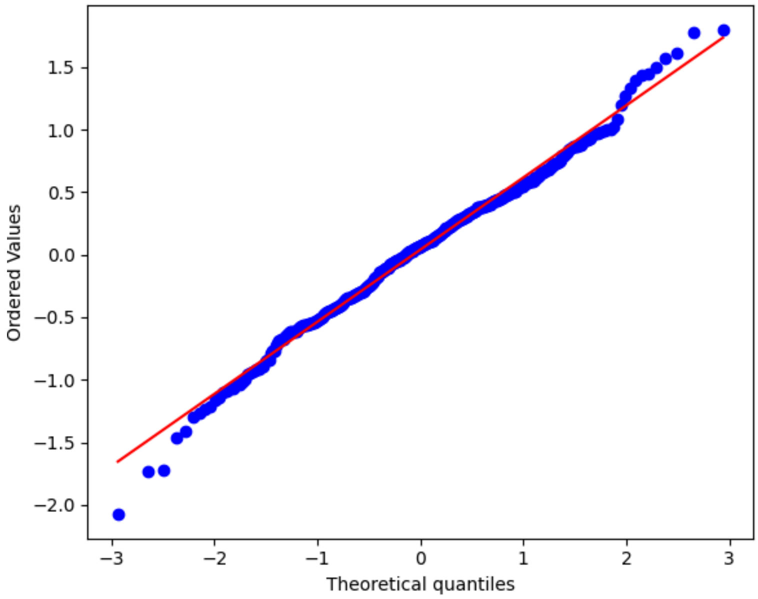


Fig. 19-c: *QQ plot of residuals*

4. Gradient Boosting

a. Training

Gradient boosting is a powerful machine learning model that can be used in both regression and classification problems. This model is based on building decision trees sequentially, where each new tree will focus on correcting the errors made by the previous ones. This iterative process allows reducing the overall prediction error and enhances the overall predictive accuracy.

To fully understand and interpret the performance of gradient boosting models, it is crucial to examine the importance of individual features. Conversely to simpler models like linear regression shown previously, the complex nature of gradient boosting makes it essential to assess which variables most significantly influence the predictions. By plotting features' importance, we can gain valuable insights into how each variable contributes to the model's decisions and understand their relative impact on the prediction outcomes.

Figure 20 below shows the top 5 most important features of our gradient boosting model. Consistently with the heatmap used during the pre-processing step (see ¶ *Relationship of the happiness score to external factors),* logged GDP per capita and healthy life expectancy emerge as the two factors with the greatest influence on the model.

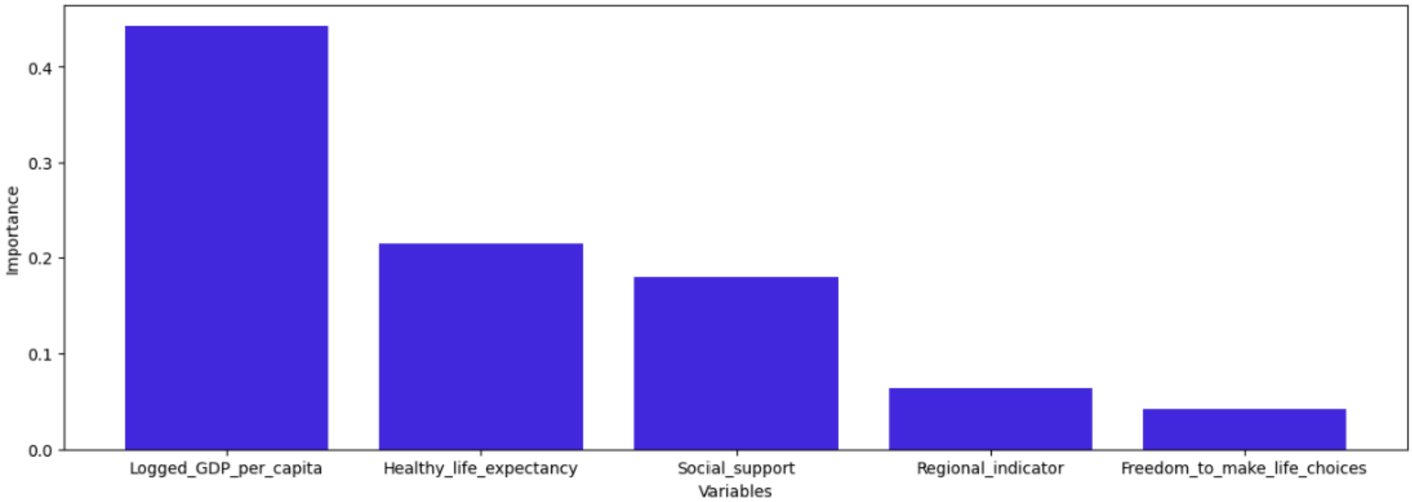


Fig.20: *Top 5 features of importance (Gradient Boosting).*

Having identified the most important features driving the predictions of our gradient boosting model, it is now essential to evaluate the model's overall performance. To do so, we used the coefficient of determination R². The following R² values were obtained:

- R² value of the train set : 0.90
- R² value of the test set: 0.84

The high values observed here suggest that the model fits the data very well. The slightly lower R² value of the test compared to the training suggests that there might be some overfitting, however not significant. Despite this slight drop in performance from training to test, the model seems to generalize reasonably well.

b. Prediction

The performance metrics we obtained after performing prediction with this model are as follows:

	Train dataset	**Test dataset**
MAE	0.27	0.34
MSE	0.12	0.20
RMSE	0.35	0.45

We can see the values are closer to 0 compared to the previous linear regression model, suggesting a better performance with fewer errors and improved predictive accuracy.
The observed increase in errors from the training set to the test set indicates some degree of overfitting, but this increase is not substantial, suggesting that the model still generalizes reasonably well to unseen data.
Overall, the performance metrics indicate that the model is making relatively accurate predictions.This is further confirmed visually by plotting the predicted values against the target values (fig. 21).

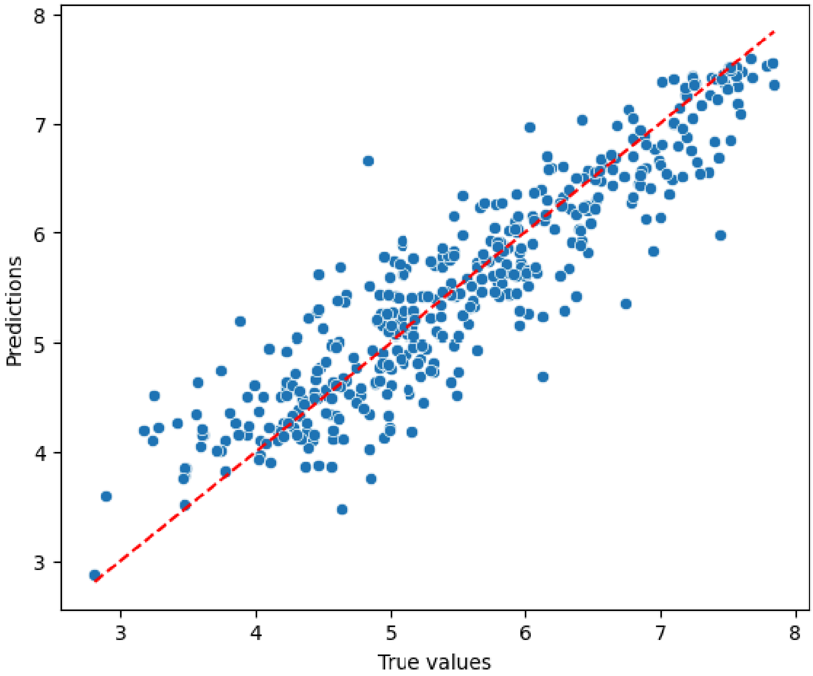


Fig.21: Comparison of actual values vs predicted values.

Once again, the residual plots are consistent with this assumption, showing patterns indicating that the model's predictions are reasonably accurate. Typically, errors randomly distributed around the baseline on the scatter plot (fig. 22-a), normal distribution on the histogram (fig. 22-b) and residuals aligning with the 45-degree line on the QQ plot (fig. 22-c).

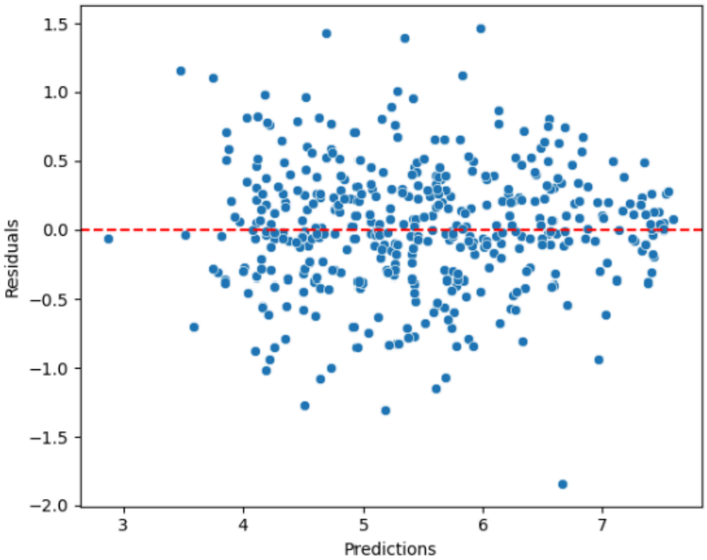


Fig. 22-a: Scatter plot of residuals.

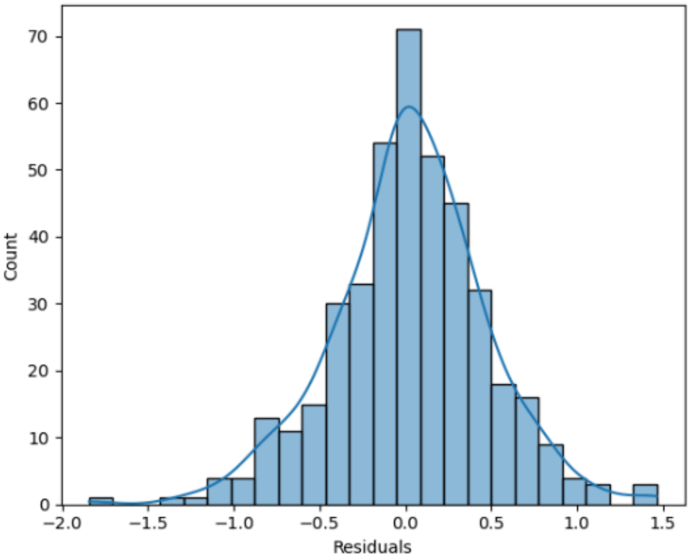


Fig. 22-b: Histogram of residuals.

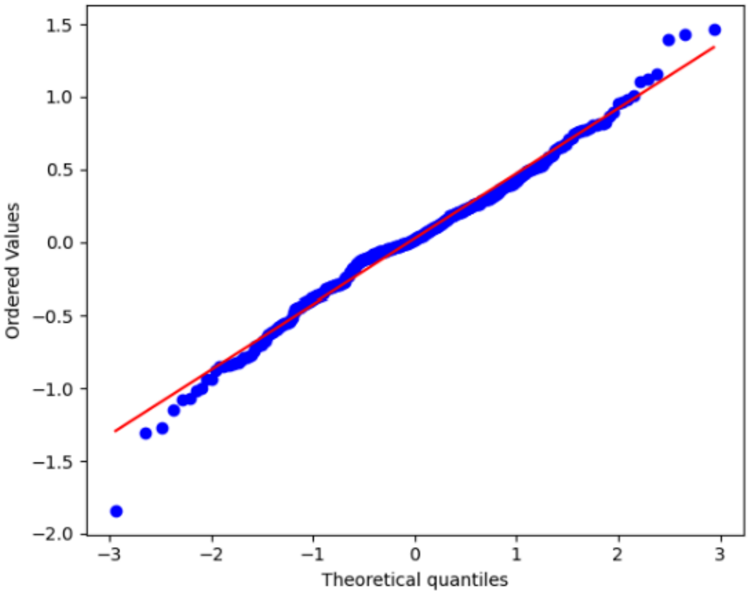


Fig. 22-c: QQ plot of residuals.

Conclusion

This table presents a comprehensive comparison of the scores and metrics generated by our models. It includes the R² scores, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) for both the training and test set.

Comparison of scores and metrics for all models

	Linear Regression	Decision Tree	Random Forest Tree	Gradient Boosting
R² train	0.75	0.74	0.77	0.90
R² test	0.73	0.69	0.73	0.84
MAE (train)	0.43	0.44	0.41	0.27
MAE (test)	0.45	0.47	0.45	0.34
MSE (train)	0.31	0.31	0.28	0.12
MSE (test	0.56	0.37	0.33	0.20
RMSE (train)	0.33	0.56	0.53	0.35
RMSE (test)	0.58	0.60	0.57	0.45

R² Scores: As observed, the Gradient Boosting model has the highest R² scores, indicating it explains the most variance in the data. Although this model shows the largest difference between training and test R² scores, the gap is not significantly greater than those of the other models, suggesting it generalizes well. A high R² score means that the model's predictions closely match the actual data, which is crucial for reliable analysis.

Mean Absolute Error (MAE): The MAE measures the average absolute difference between predicted and actual values. It is straightforward to interpret and less sensitive to outliers, making it a robust metric for model evaluation. Our Gradient Boosting model excels here, with a lower MAE than the other models, indicating more accurate predictions on average. This suggests that the model consistently makes smaller errors in its predictions.

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE): These metrics penalize larger errors more than MAE, providing a more stringent measure of model performance. The Gradient Boosting model achieves the lowest scores in both MSE and RMSE among all our models, further demonstrating its superior accuracy. Lower MSE and RMSE values indicate that the model not only makes fewer errors but also avoids large prediction errors, which is essential for maintaining high prediction quality.

Overall, these metrics collectively show that our Gradient Boosting model is the most accurate and reliable for predicting the Ladder Score. Its superior performance highlights its robustness and effectiveness in capturing complex relationships and interactions between features, making it well-suited for the multidimensional nature of happiness data. It also underscores the complexity of happiness as a global phenomenon.

Happiness is influenced by a lot of variables, including economic indicators, social support, life expectancy, freedom, generosity and perceptions of corruption, among others. Using advanced machine learning techniques like gradient boosting is crucial for capturing these intricate patterns. It allows informing institutions and researchers having the purpose of developing comprehensive strategies, to improve well-being worldwide.

With additional time allocated to this project, refinements could be performed on the different models, such as optimizing hyperparameters or incorporating additional data. This would allow improving and validating their relative robustness across different datasets, and ensuring their reliability and applicability.

Finally, as an opening focus, further investigations could be conducted towards the regional indicator, which we observed was consistently among the top 5 most important features. This intriguing observation suggests a need for geographical analysis to identify patterns and discrepancies in happiness levels across different continents and countries. Additionally, exploring the socio-economic and cultural contexts that may explain these regional differences would provide deeper insights.

Appendix

Data Sources

1. Datasets were extracted from Kaggle: World Happiness Report 2021 (kaggle.com)
2. Cantril Self-Anchoring Scale: Understanding How Gallup Uses the Cantril Scale

Notebooks

- Data exploration of df_pre2021 and df_2021 datasets: 01_exploration.ipynb
- Data visualization: 02_DataViz.ipynb
- Data preprocessing including data merge of df_pre2021 and df_2021: 03_preProcessing.ipynb
- Data modeling: 04_modelisation.ipynb
- Report cover page design: 05_CoverPageDesign.ipynb