Contents

I. Introduction

The World Happiness Report has increasingly drawn global attention in recent years, as it provides valuable insights into the well - being of populations across different countries. Happiness, a complex and multi - faceted concept, is influenced by a variety of factors such as economic stability, social support, health conditions, and freedom of choice.

As we enter the mid - 2020s, understanding the trends and changes in the world's happiness levels becomes crucial. The question then arises: How have the happiness scores of countries evolved from 2024 to 2025? Are there consistent patterns among high - ranking and low - ranking countries? By analyzing the World Happiness Index maps and rankings for 2024 and 2025, we aim to explore these questions, uncover the underlying factors contributing to the differences in happiness scores, and gain a deeper understanding of the global landscape of well - being.

II. Data Preprocessing

2.1 Cleaning

There is already a happiness index and related dataset from 2018 to 2024. Considering the completeness of the data and its impact on overall statistics, the mean is used to fill in missing values.

2.2 Transformation

The data for 2022 has formatting issues (such as commas in numbers), excluding non numeric columns (such as Country and Happiness Rank), and only converting numeric columns to object types that contain numerical values.

2.3 Feature Selection

Calculate the correlation coefficient between each feature and the happiness index score. Features with higher correlation may be more important for predicting happiness index scores.

Firstly, Shapiro Wilk normality test is performed on the features of each dataset, and the Spearman correlation coefficient is selected for non normal distributions.

Then a correlation threshold (0.2) was set, retaining only features with a correlation higher than this threshold with the happiness index score, and ultimately excluding the Genealogy column.

2.4 Data Combination

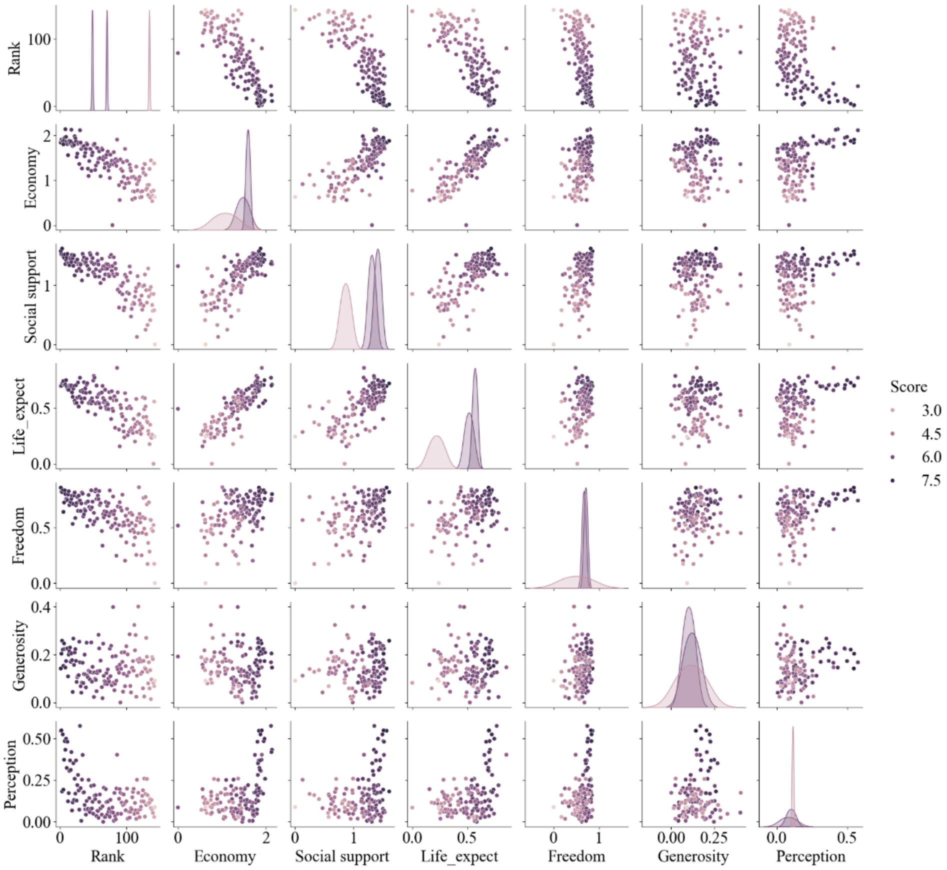
Add a Year column to each dataset to indicate the year to which the data belongs. Merge all datasets into one combined.df for ease of subsequent processing.

2.5 Data Standardization

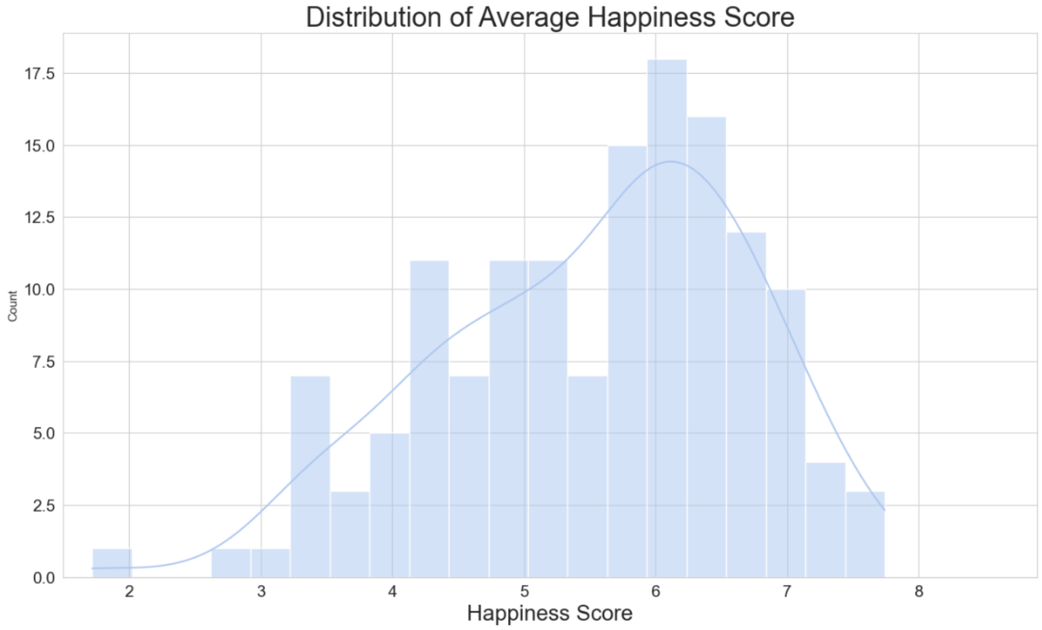
Using StandardScaler to standardize feature data, converting the standardized data into a DataFrame format, so that different features have the same scale, avoiding certain features from having excessive impact on model training due to large numerical ranges, and helping to improve model training effectiveness and convergence speed.

III. Data Visualization

Take 2024.xlsx as an Example.

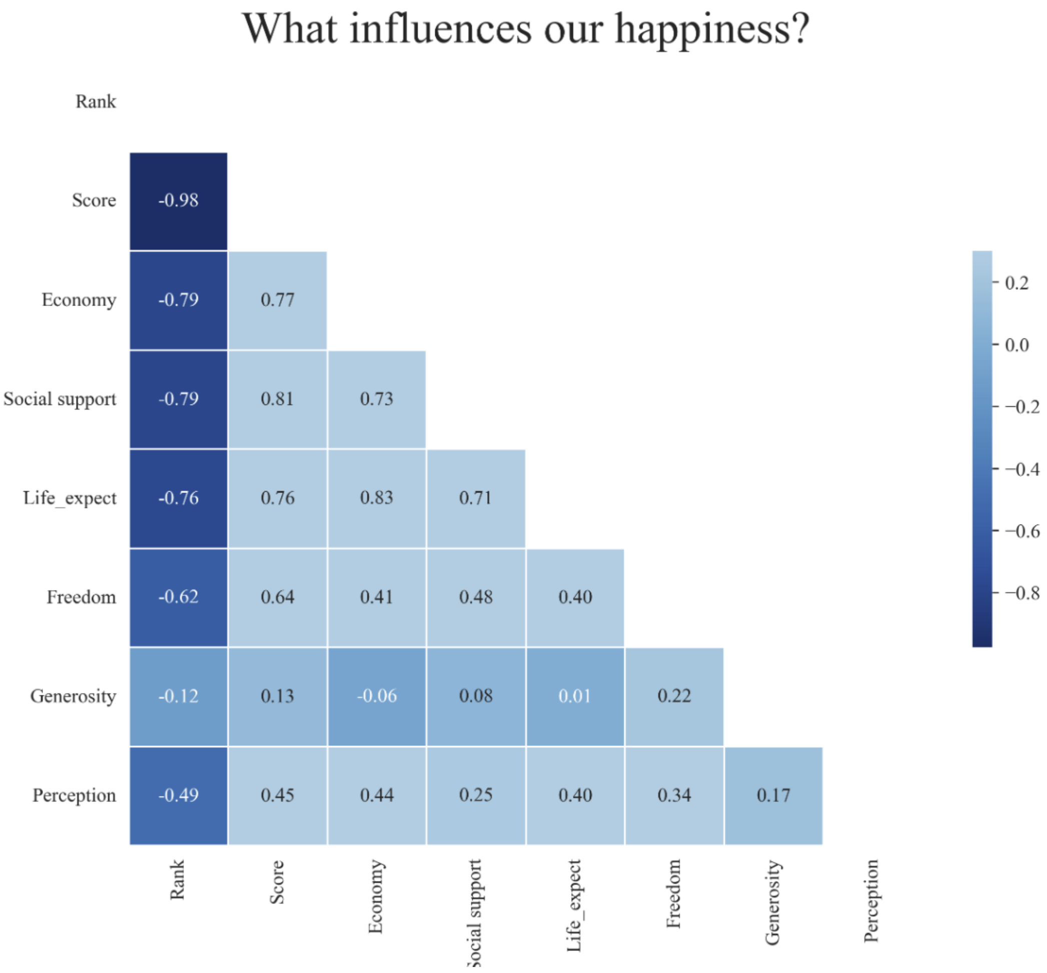


From this paired relationship diagram, it can be seen that there is rich correlation between the variables. The rank of happiness is negatively correlated with variables such as economy and social support, indicating that countries with favorable economic conditions and good social support rank higher. There is a positive correlation between economic and social support, as well as healthy life expectancy, reflecting the supportive role of the economy in various aspects of social life. Social support is closely positively correlated with healthy life expectancy and happiness score, highlighting its importance for residents' health and happiness. The correlation between Freedom and other variables may not be intuitive, but it is traceable. Generosity has weak correlation with other variables. Perception of corruption is negatively correlated with rankings and happiness scores, indicating that perception of corruption is a negative factor affecting happiness. Meanwhile, the distribution differences of points with different happiness scores in the graph further confirm the influence of variable relationships on happiness.

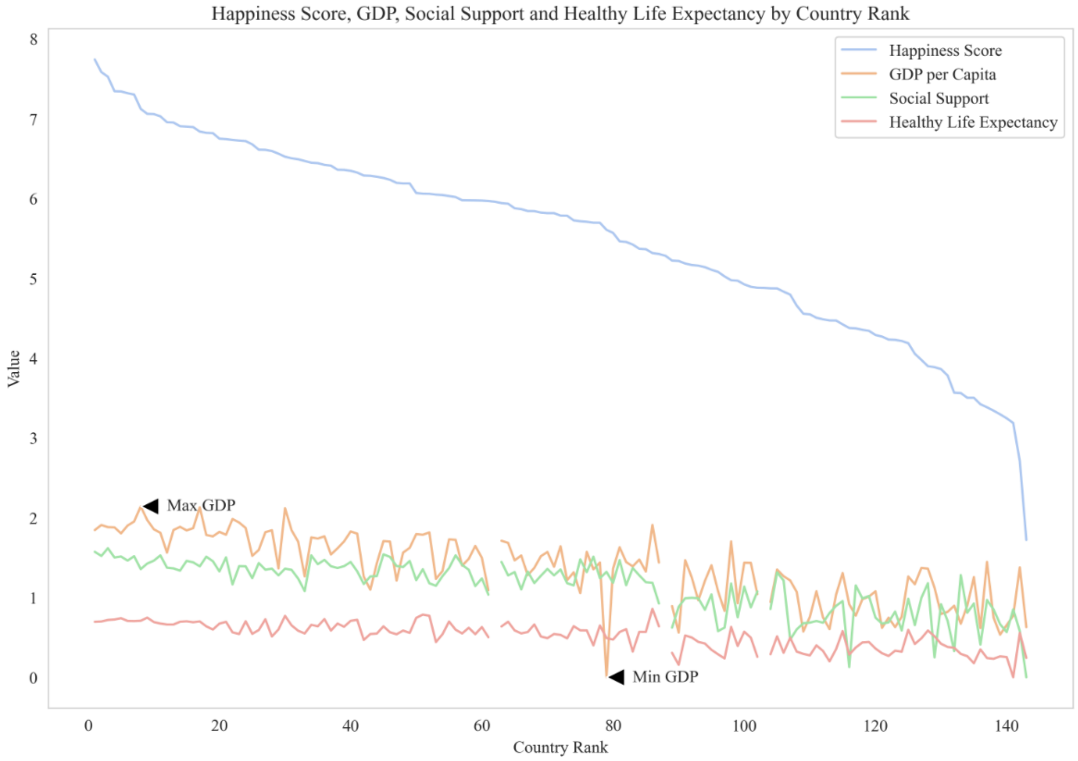


The statistical chart of Distribution of Average Happiness Score combines histogram and kernel density estimation curve. The histogram displays the frequency of different happiness score intervals in a bar shape, and the kernel density estimation curve (blue) shows the distribution pattern of the data. From the graph, it can be seen that happiness scores are concentrated between 4-7 points, with the highest frequency around 6 points. The kernel density curve shows a single peak and is approximately symmetrical, and the overall distribution is approximately normal. The frequency is extremely low around 2 and 8 points, indicating that there are very few extreme happiness scores. This indicates that the average happiness score is concentrated around

6 points, with very few extremely low or high happiness scores, and the distribution is approximately normal.



The correlation coefficient heatmap shows the correlation between variables such as Rank (happiness ranking), Score (happiness score), Economy (economy), Social support (social support), Life-expect (healthy life expectancy), Freedom (degree of freedom), Generosity (generosity), and Perception (perception of corruption). The depth of color represents the magnitude of the correlation coefficient, with darker blue indicating stronger negative correlation and darker light blue indicating stronger positive correlation. It can be seen that Rank is strongly negatively correlated with most other variables, indicating that countries ranked high in happiness tend to perform better in terms of economic and social support; The correlation between Score and various variables reflects that economic, social support, and healthy life expectancy have a significant positive impact on happiness scores; The correlation between genetics and other variables is generally weak; Perception is negatively correlated with other variables, indicating that the stronger the perception of corruption, the worse the performance in other aspects may be. This indicates that economic, social support, and healthy life expectancy are important positive factors that affect happiness, while the perception of corruption is a negative factor, and generosity has a relatively small impact on happiness.



Happiness Score, GDP, Social Support and Healthy Life Expectancy by Country Rank The line chart shows the changes in Happiness Score, GDP per Capita, Social Support, and Healthy Life Expectancy values for different country rankings. The blue line in the figure represents the happiness score, which shows an overall downward trend, indicating that as the country ranks lower, the happiness score gradually decreases; The orange line represents per capita GDP, with fluctuations indicating ”Max GDP” (highest GDP) and ”Min GDP” (lowest GDP); The green line represents social support, while the red line represents healthy life expectancy, both of which fluctuate. This indicates that there is a negative correlation between happiness scores and national rankings, and that per capita GDP, social support, and healthy life expectancy have different performances under different national rankings, reflecting to some extent the comprehensive impact of these factors on happiness scores.

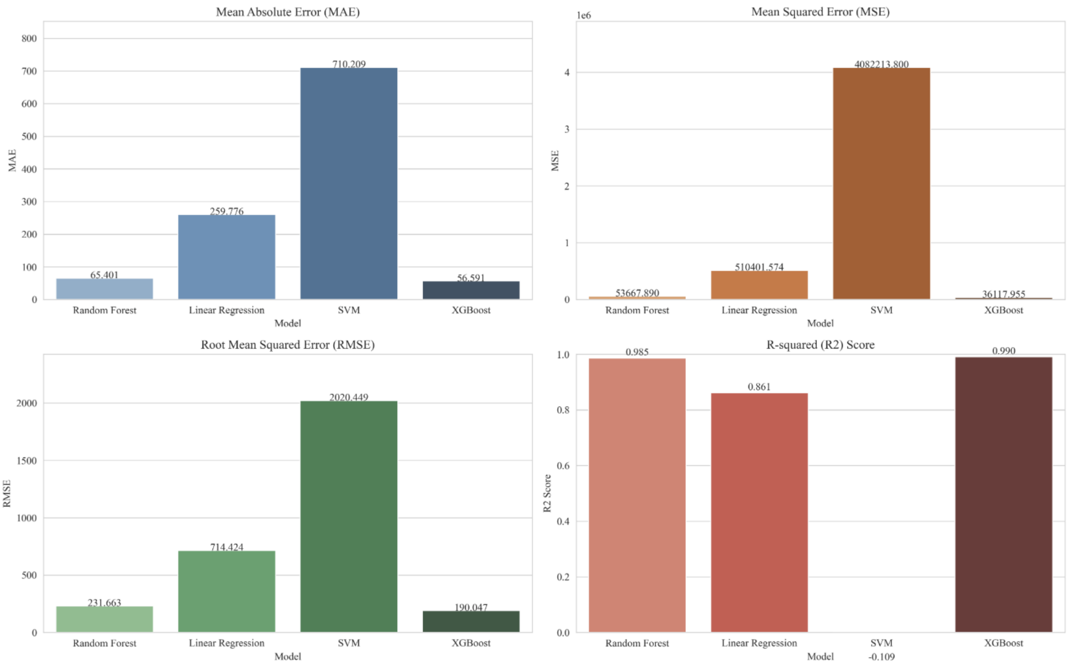
IV. Develop a predictive model

4.1 Model Preparation

4.1.1 *Characteristic and target variable definitions*

Economy, Social support, Healthy life expectancy, Freedom, Perceptions of corruption, and Year are selected as characteristic variables X, which are considered to have an impact on Happiness Score. And Happiness Score itself serves as the target variable y.

4.1.2 *Model Comparasion*



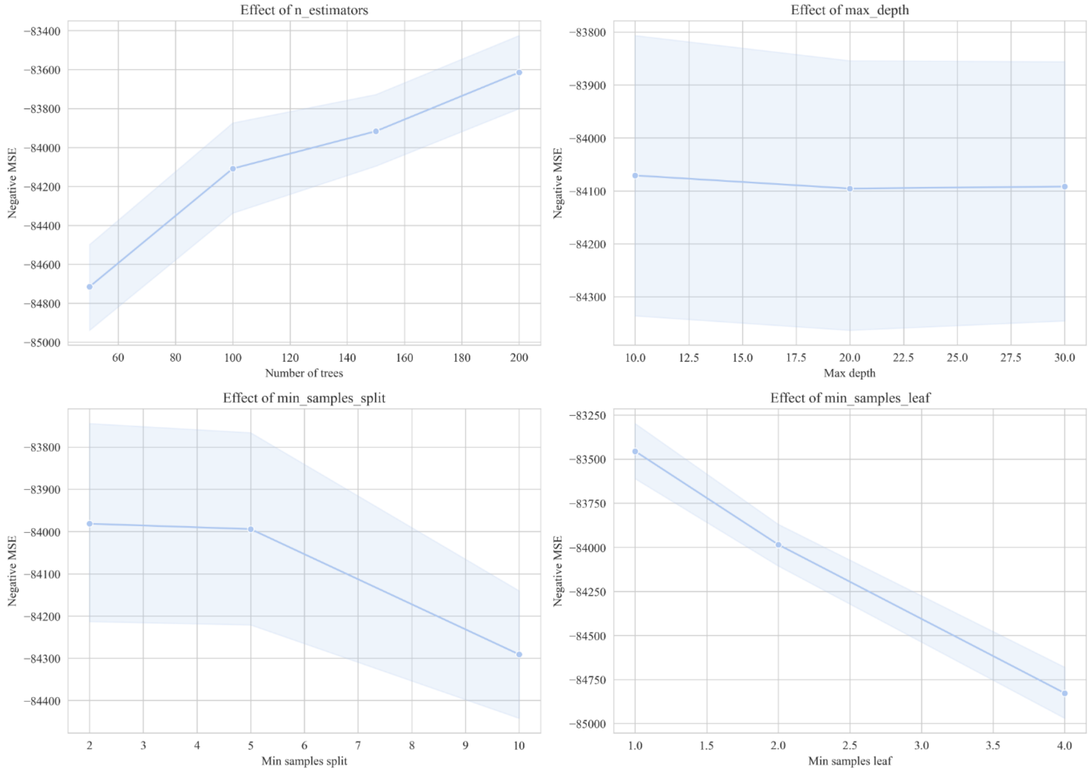
From the comprehensive analysis of these four charts (MAE, MSE, RMSE, R2 Score), it can be seen that when selecting a model, it is usually desirable to have smaller values for MAE, MSE, and RMSE (indicating smaller prediction errors) and larger values for R2 Score (indicating a higher degree of fit of the model to the data). Due to the small error and high fitting degree of the random forest model, combined with the complexity and generalization ability of the model, the random forest model is selected for prediction.

4.1.3 *Divide the training set and testing set*

Divide the standardized feature data and target variables into training sets X_train, y_train, and testing sets X_test in an 80:20 ratio y_test Set random_state=42 during partitioning to ensure consistency of partitioning results every time the code is run, facilitating model evaluation and comparison.

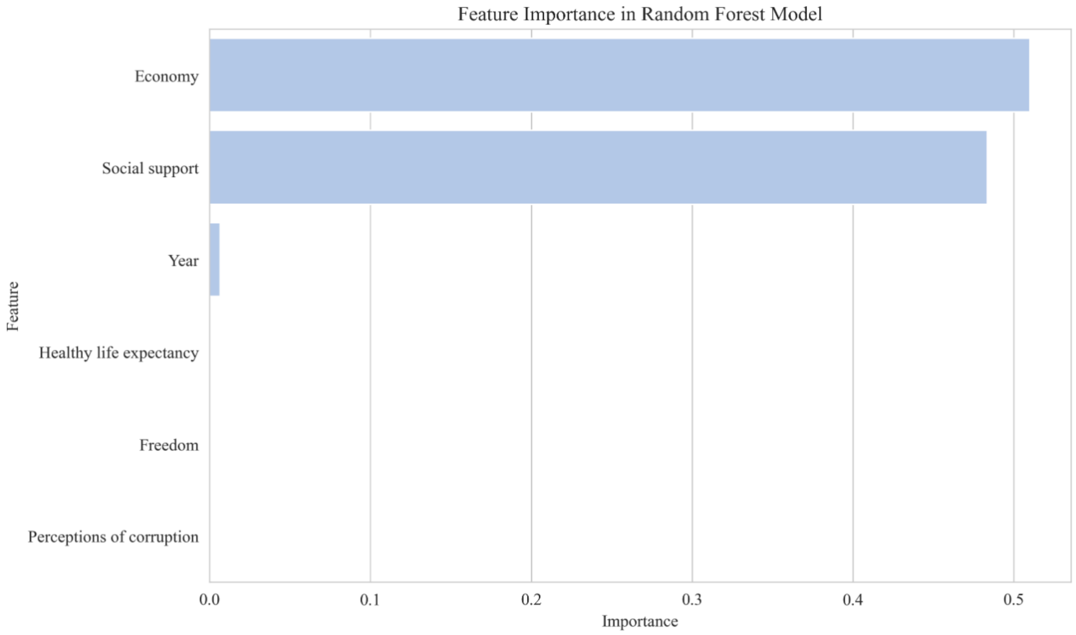
4.2 Model Training

4.2.1 *Parameter tuning*



Perform 5-fold cross validation using grid search to evaluate 144 different parameter combinations (candidate parameters), requiring a total of 720 model fits. “Best parameters: ’max$_{d}epth' : 10,' min_{s}amples_{l}eaf' : 1,' min_{s}amples_{s}plit' : 2,' n_{e}stimators' :$ 200”

Best cross validation score: 82527.4465 ”: The cross validation score here is the opposite of the negative mean square error (because Neg’mean_Squared_error is used as the evaluation metric). The higher the value, the smaller the prediction error of the model on the cross validation set and the better its performance. This score indicates the performance of the model under the optimal parameter combination in cross validation.



Economy and social support are the two most important factors for predicting happiness scores, and their feature importance scores are significantly higher than other features, indicating that in this model, economic status and social support have the greatest impact on happiness scores. The feature importance score of Year is extremely low, indicating that this factor has almost no effect in predicting happiness scores. The relatively low importance scores of Healthy Life Expectancy, Freedom, and Perceptions of Corruption indicate that their impact on predicting happiness scores is relatively small, but not completely ineffective.

4.3 Evaluation

Predict y_pred on the test set X_test, and then calculate the mean squared error (MSE) using mean_Squared_error, which measures the average squared error between the predicted and true values. The smaller the value, the better the model's prediction performance; Use r2_store to calculate the R-squared value, which represents the goodness of fit of the model to the data. The range of values is between 0 and 1, and the closer it is to 1, the better the fitting effect of the model. Print these two evaluation metrics to understand the performance of the model.

$\vartriangle$ MSE

Mean square error is the average of the squared differences between predicted and true values, which measures the average degree of deviation between the predicted and true values. The calculation formula is as follows:

Let $y_i$be the true value$\hat{y}_i$is the predicted value$Ifn$is the number of samples, the formula for calculating the mean square error is expressed as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

△R^2

$R^2$is used to measure the goodness of fit of a regression model, which represents the proportion of variance that the model can explain. The closer its value is to 1, the better the fitting effect of the model on the data.

Let $bary$be the average of the true values$Y_i$is the true value$\hat{y}_i$is the predicted value .If n is the sample size, then the calculation formula for R $\hat{2}$ is expressed as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Among them, molecule $sm_{i=1}^{n}(y_i- hat)y_i^2$represents the Sum of Squared Residuals (SSR), which is the sum of squared differences between the predicted values and the true values of the model; The denominator $sm_{i=1}^{n}(y_i - bary)^2$represents the Total Sum of Squares (SST), which is the sum of squared differences between the true value and the mean of the true value.

4.4 Result

Use the trained model to predict the data for 2025 and obtain the predicted Happiness Score for each country in 2025. The results are as follows

Below are the global distribution maps of happiness index for 2024 and 2025.

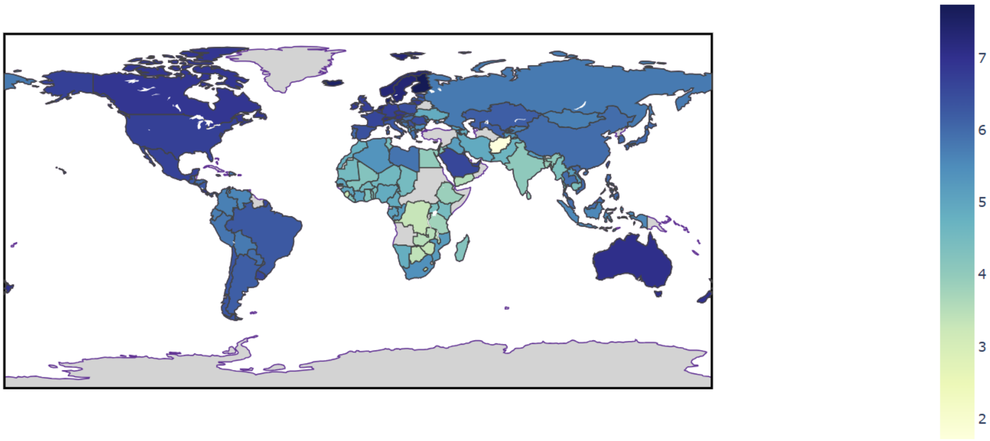


Figure 1 World Happiness Index Map in 2024

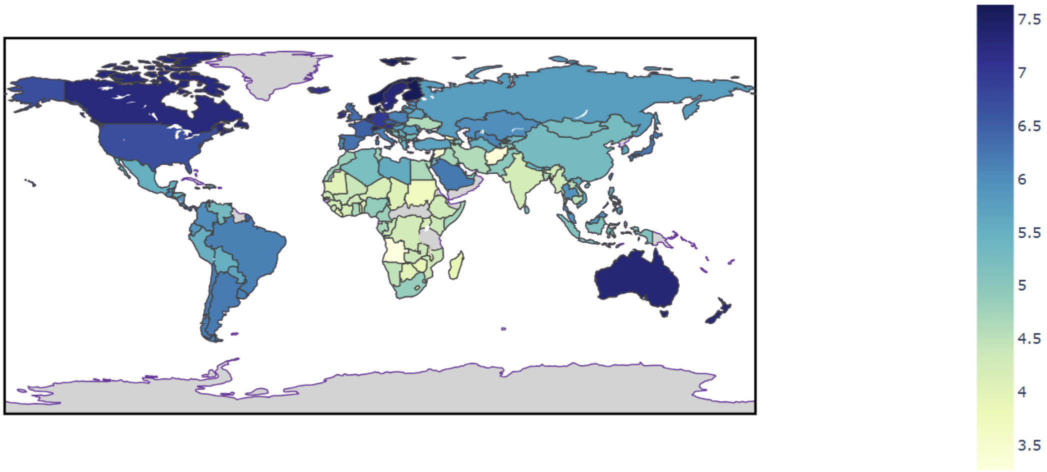


Figure 2 World Happiness Index Map in 2025

From the 2024 and 2025 World Happiness Index maps, it can be seen that the happiness index of Nordic and some Western countries such as Finland, Norway, Canada, and the United States has been at a high level for a long time, presented in dark blue to blue-green; Some African countries are mostly light green to light yellow in color, with low happiness index; There are significant differences among Asian countries. At the same time, compared with the two-year map, some countries have changed colors and their indices have fluctuated. Although the overall distribution pattern is generally stable, local dynamic changes reflect the complex and continuous changes in factors affecting the happiness index. Countries need to constantly adjust their development strategies to ensure or enhance residents' sense of happiness.

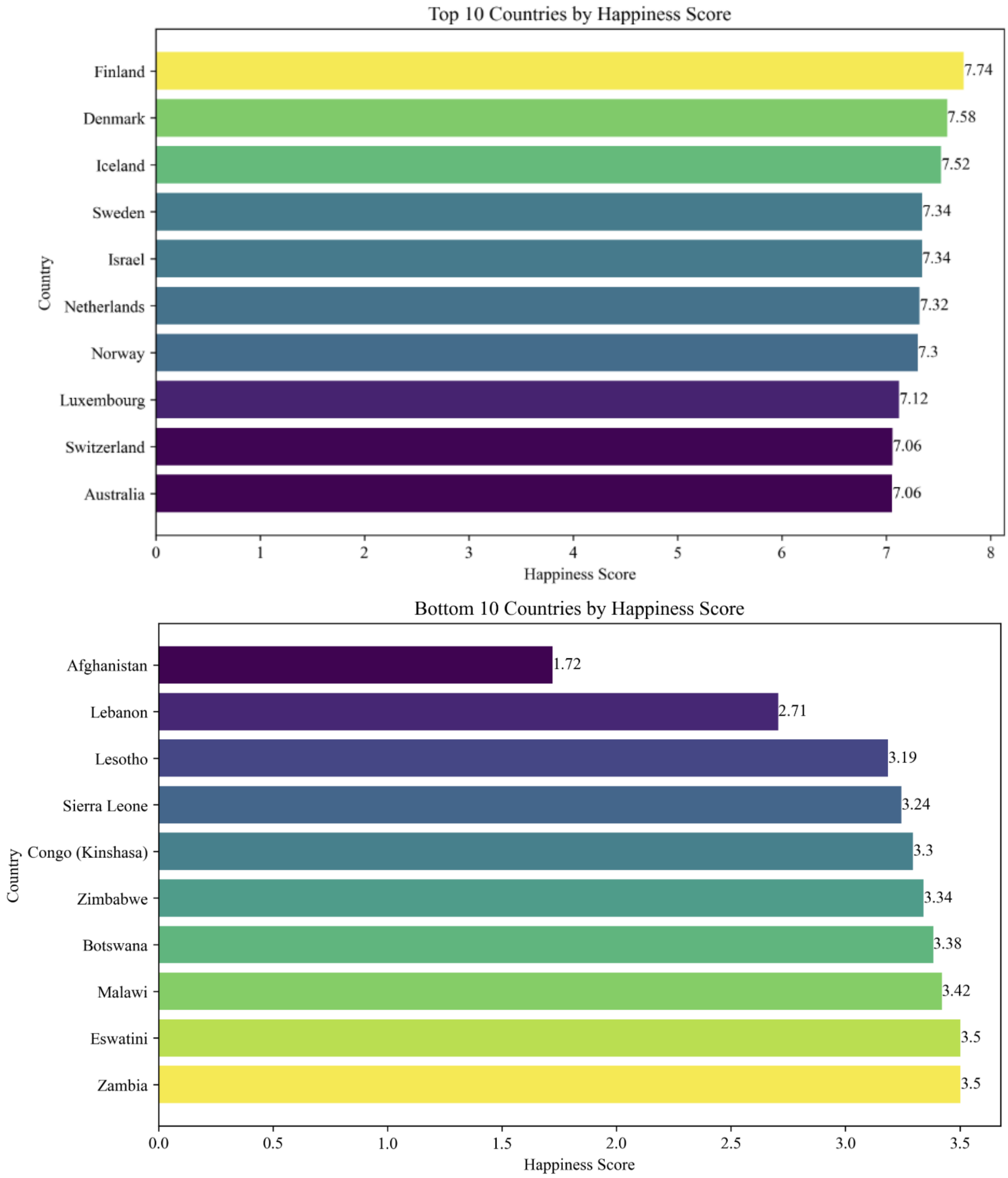


Figure 3 Happiness Rank in 2024

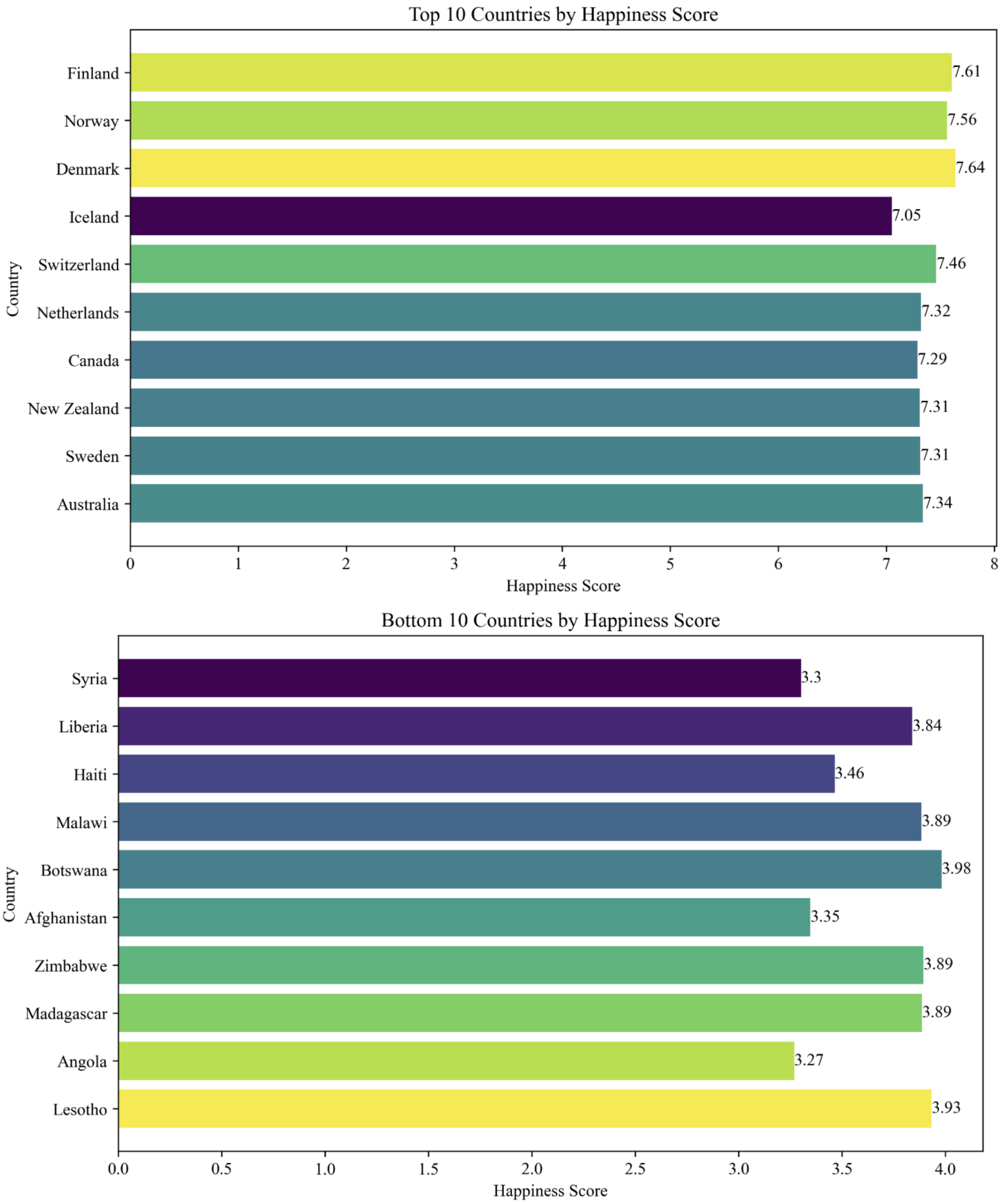


Figure 4 Happiness Rank in 2025

According to the ranking charts of happiness index in 2024 and 2025, Nordic countries such as Finland, Norway, and Denmark have consistently ranked among the top in happiness index in these two years, with high happiness scores, demonstrating their long-term advantages in improving people's happiness; Some Western countries such as Switzerland, the Netherlands, Australia, etc. are also stable in the high segment. However, countries such as Afghanistan and Syria have been in the low segment of happiness index for two years, reflecting that these regions are facing significant difficulties in social, economic, and livelihood aspects, and the task of improving people's happiness is arduous. Overall, there are significant differences in happiness indices among different countries, with high happiness index countries mostly located in economically developed regions with well-established social welfare systems, while low happiness index countries are constrained by various unfavorable factors.

V. Conclusions

According to the analysis of the predicted 2025 World Happiness Index chart and table, Nordic countries such as Finland, Norway, and Denmark rank high in happiness index, with scores mostly around 7.5 or above, while countries such as Syria and Afghanistan rank low, with scores around 3.3-3.4. The happiness score is concentrated between 3-7.6, with a larger number of countries scoring 4-6. From a regional perspective, European countries generally perform well, while Asian countries have significant differences, and African countries generally have low scores. This may be related to the situation of various countries in terms of economic development, social support, healthy life expectancy, perception of freedom and corruption. Countries with high happiness indices often perform well in these areas, while those with low happiness indices may have shortcomings.

VI. References

[1] A. Saffari, C. Leistner, J. Santner, M. Godec and H. Bischof, ”On-line Random Forests,” 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, Kyoto, Japan, 2009, pp. 1393-1400, doi: 10.1109/ICCVW.2009.5457447. keywords: Machine learning;Application software;Usability;Machine learning algorithms;Computer vision;Computational efficiency;Training data;Bagging;Decision trees;Radio frequency