Detailed Analysis of World Happiness Report (2015-2019)

Solomon Silwal

February 22, 2025

Contents

1 Introduction

Happiness is a crucial indicator of societal well-being and development. The World Happiness Report measures happiness across countries based on key social and economic indicators. This project aims to analyze happiness scores from 2015 to 2019 using machine learning models to predict happiness and identify key contributing factors.

2 Data Preprocessing

2.1 Dataset Overview

The dataset includes happiness scores and socio-economic indicators such as GDP per capita, social support, life expectancy, and governance factors. The data is sourced from the World Happiness Report and includes records for multiple years.

2.2 Data Cleaning and Handling Missing Values

- Merged datasets from 2015 to 2019, ensuring column consistency.
- Dropped irrelevant features such as 'Dystopia Residual'.
- Handled missing values by applying mean imputation where necessary.
- Standardized column names across years to maintain consistency.

2.3 Feature Selection and Normalization

- Removed low-impact features such as 'Generosity', 'Year', and 'Happiness Rank'.
- Standardized features using Z-score normalization for regression models.
- Applied Min-Max scaling for tree-based models to enhance performance.

3 Exploratory Data Analysis (EDA)

3.1 Statistical Summary

A statistical summary was generated to understand the distribution of key variables:

Feature	Mean	Standard Deviation	Min	Max
Happiness Score	5.38	1.13	2.69	7.77
GDP per Capita	0.92	0.41	0.00	2.10
Social Support	0.72	0.17	0.00	0.97
Life Expectancy	0.61	0.25	0.00	1.14
Freedom	0.41	0.15	0.00	0.72
Corruption	0.11	0.09	0.00	0.44

3.2 Visualizations

Multiple visualizations were generated to explore data trends:

- Histograms showing the distribution of Happiness Scores and other variables.
- Boxplots detecting outliers in the dataset.
- Scatter plots highlighting relationships between GDP per Capita and Happiness Score.
- Line charts revealing stability in happiness trends over time.

3.2.1 Histogram of Happiness Score and Related Features

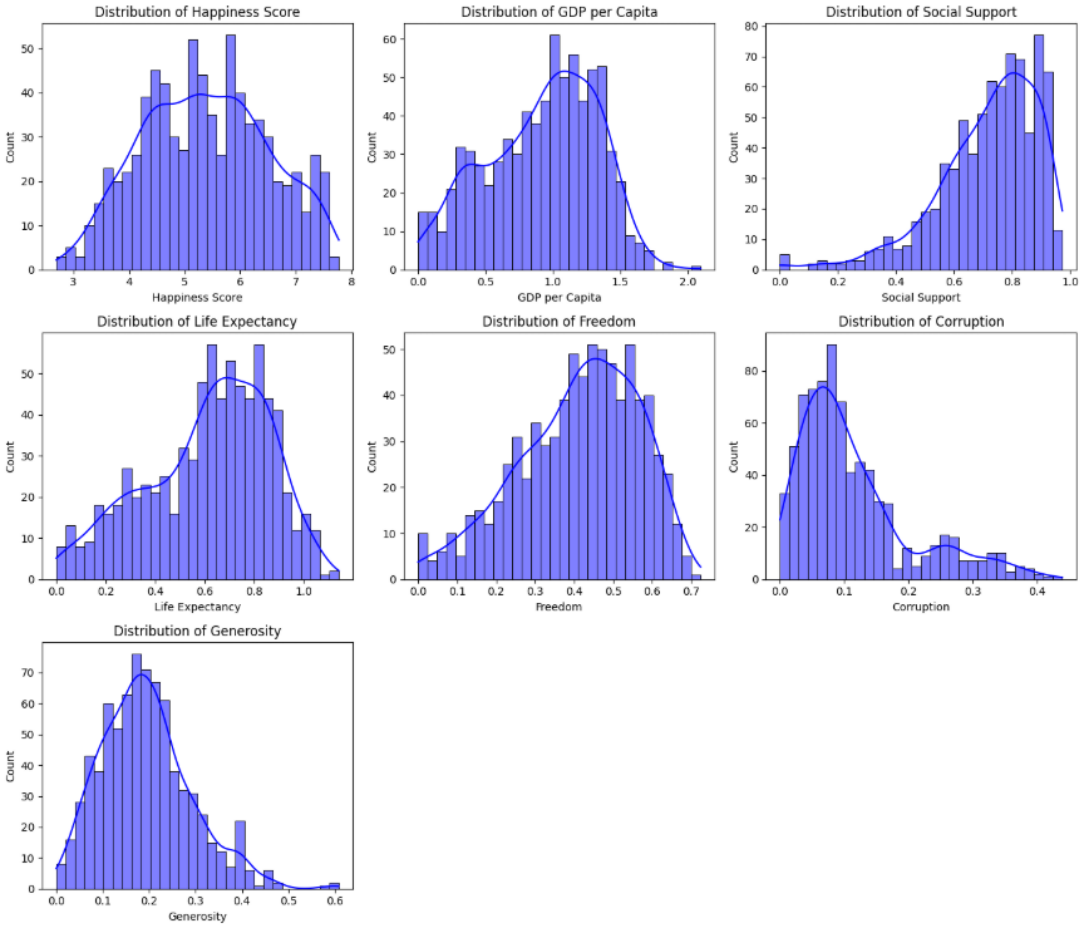

Figure 1: Histograms showing the distribution of Happiness Scores, GDP per Capita, Social Support, Life Expectancy, Freedom, Corruption, and Generosity.

Description: The histograms reveal the distribution of key features. Most distributions exhibit a right-skewed nature, suggesting an unequal spread of happiness-related factors across countries.

3.2.2 Boxplots for Outlier Detection

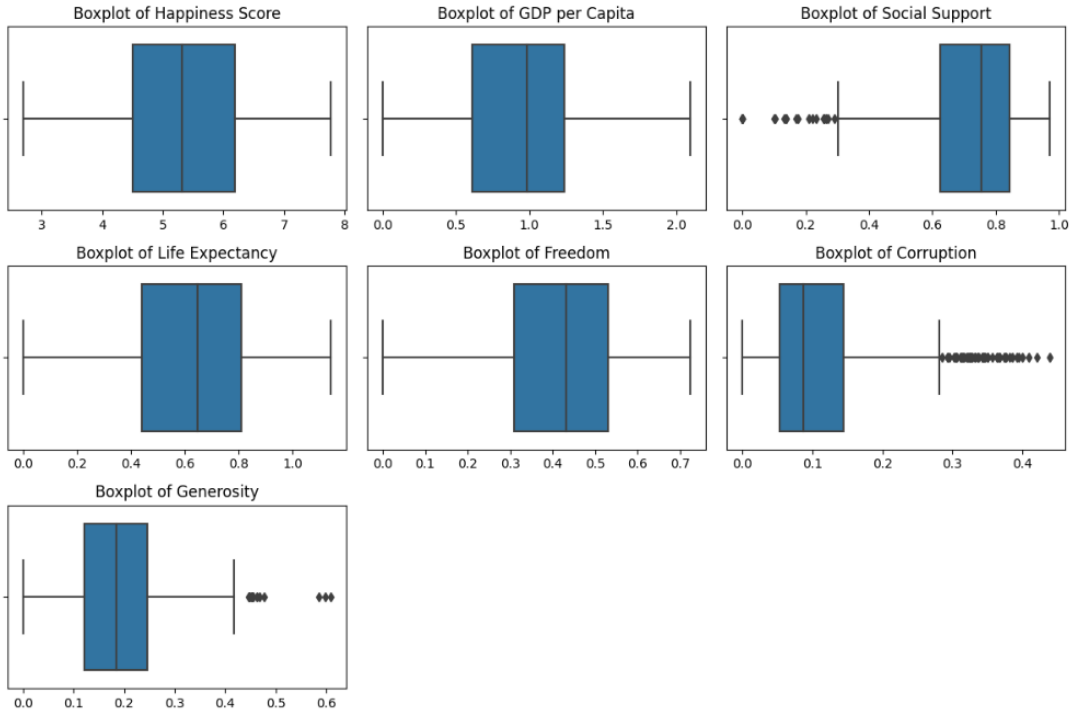

Figure 2: Boxplots identifying potential outliers in key variables.

Description: Boxplots help visualize the spread of data and detect potential outliers. Corruption and Social Support show significant outliers, indicating varying levels across different nations.

3.2.3 Correlation Heatmap

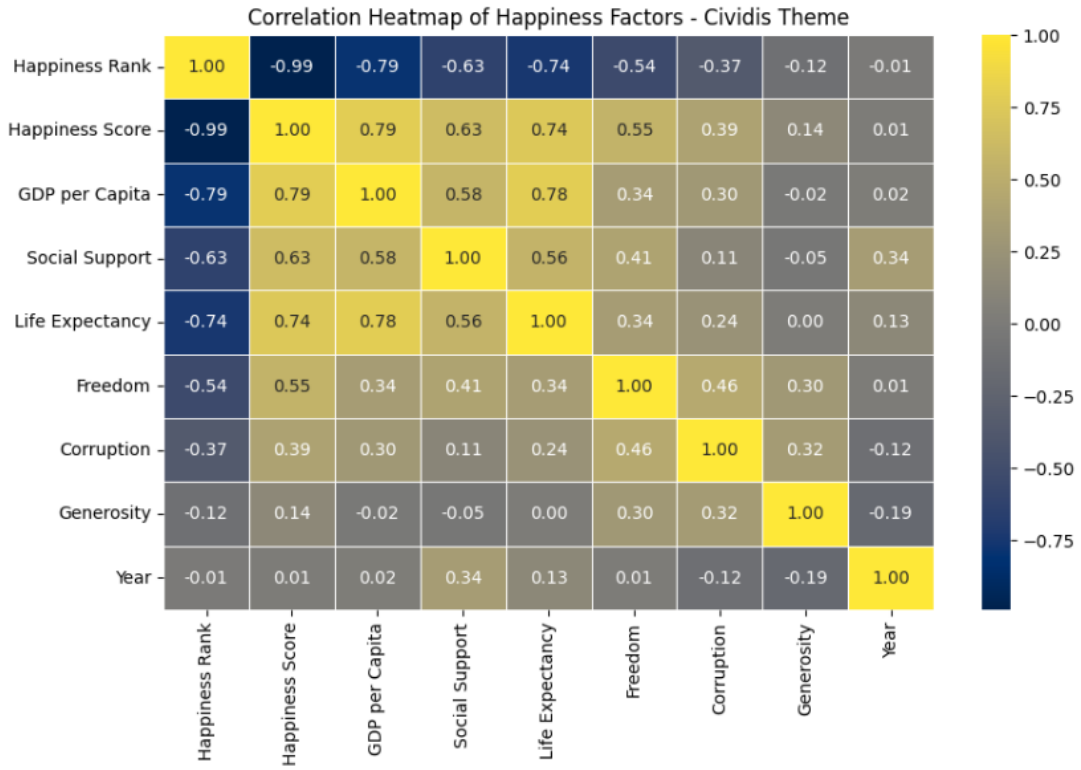

Figure 3: Correlation heatmap highlighting relationships among key happiness indicators.

Description: The heatmap reveals strong positive correlations between GDP per Capita, Life Expectancy, and Happiness Score, while Corruption shows a negative relationship.

3.2.4 Correlation Between Features and Happiness Score

Feature	Correlation with Happiness Score
GDP per Capita	0.79 (Strong)
Social Support	0.63 (Moderate-Strong)
Life Expectancy	0.74 (Strong)
Freedom	0.55 (Moderate)
Corruption	0.39 (Moderate)
Generosity	0.14 (Weak)
Year	0.01 (Not useful)
Happiness Rank	-0.99 (Redundant)

Figure 4: Correlation of Features with Happiness Score.

Description: This table highlights the correlation values of various features with the Happiness Score. GDP per Capita and Life Expectancy exhibit the strongest positive correlation, suggesting economic stability and health conditions significantly impact happiness. Social Support and Freedom also show moderate correlations, while Corruption has a weaker but noticeable negative correlation. Features like Generosity and Year have minimal impact, and Happiness Rank is redundant due to its inverse correlation with Happiness Score.

3.2.5 Scatter Plots of Key Relationships

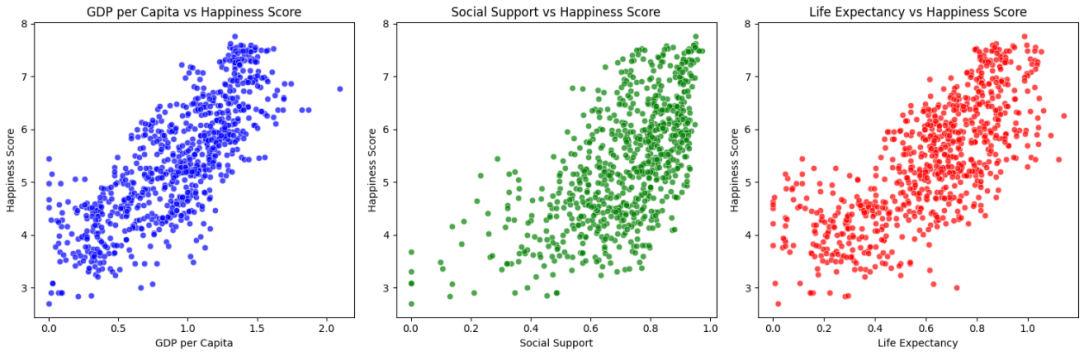

Figure 5: Scatter plots showing relationships between GDP per Capita, Social Support, Life Expectancy, and Happiness Score.

Description: The scatter plots show a strong positive trend between GDP per Capita and Happiness Score, indicating economic well-being as a crucial factor in happiness.

3.2.6 Line Charts of Trends Over Time

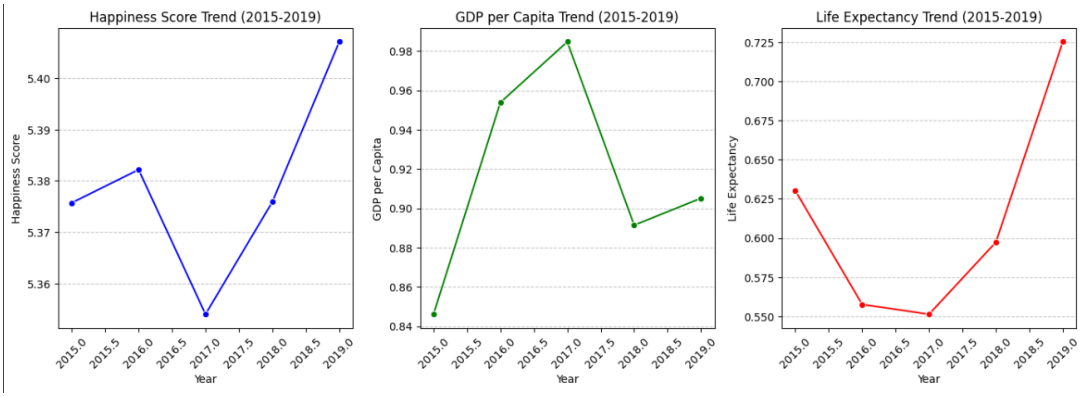

Figure 6: Line charts illustrating trends in Happiness Score, GDP per Capita, and Life Expectancy from 2015 to 2019.

Description: The line charts highlight trends over time, showing that happiness levels have remained relatively stable with minor fluctuations.

3.2.7 Happiness Score Trends for Selected Countries

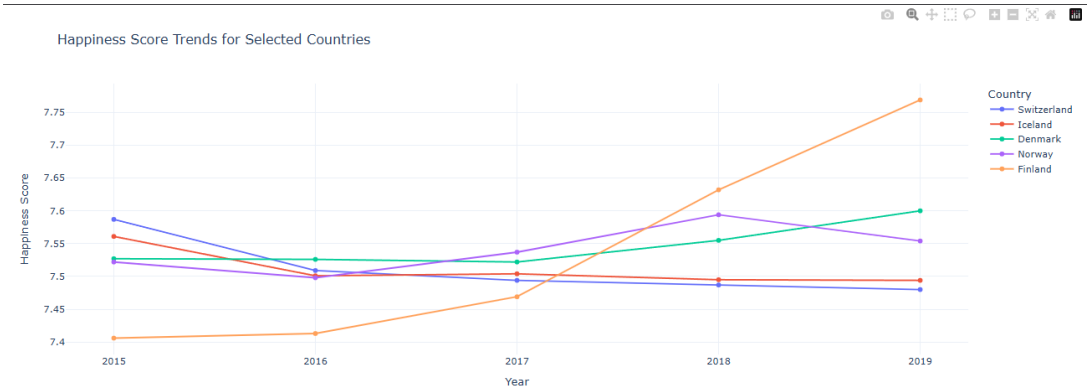

Figure 7: Happiness Score Trends for Selected Countries from 2015 to 2019.

Description: This line plot illustrates the happiness score trends for selected countries—Switzerland, Iceland, Denmark, Norway, and Finland. While most countries exhibit relatively stable trends, Finland shows a significant increase in happiness scores over time. This could indicate effective governance, economic growth, or improvements in social support systems in Finland compared to others.

3.2.8 Freedom vs Happiness Score Analysis

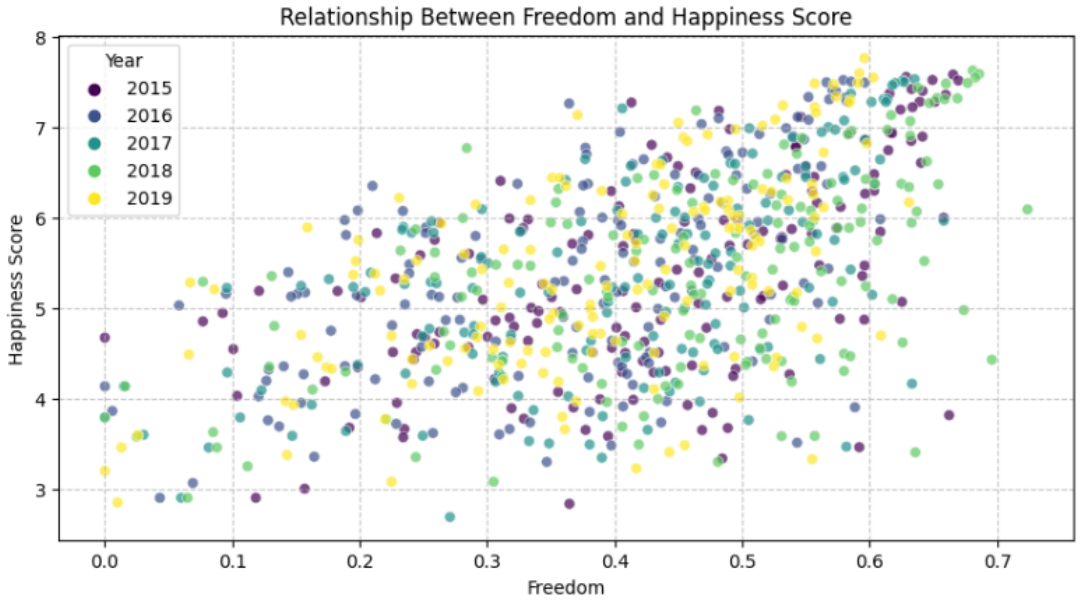

Figure 8: Scatter plot showing the relationship between Freedom and Happiness Score.

Description: Countries with higher levels of personal freedom tend to have higher happiness scores, as indicated by the positive trend in the scatter plot.

3.2.9 Corruption Perception Trends Over Time

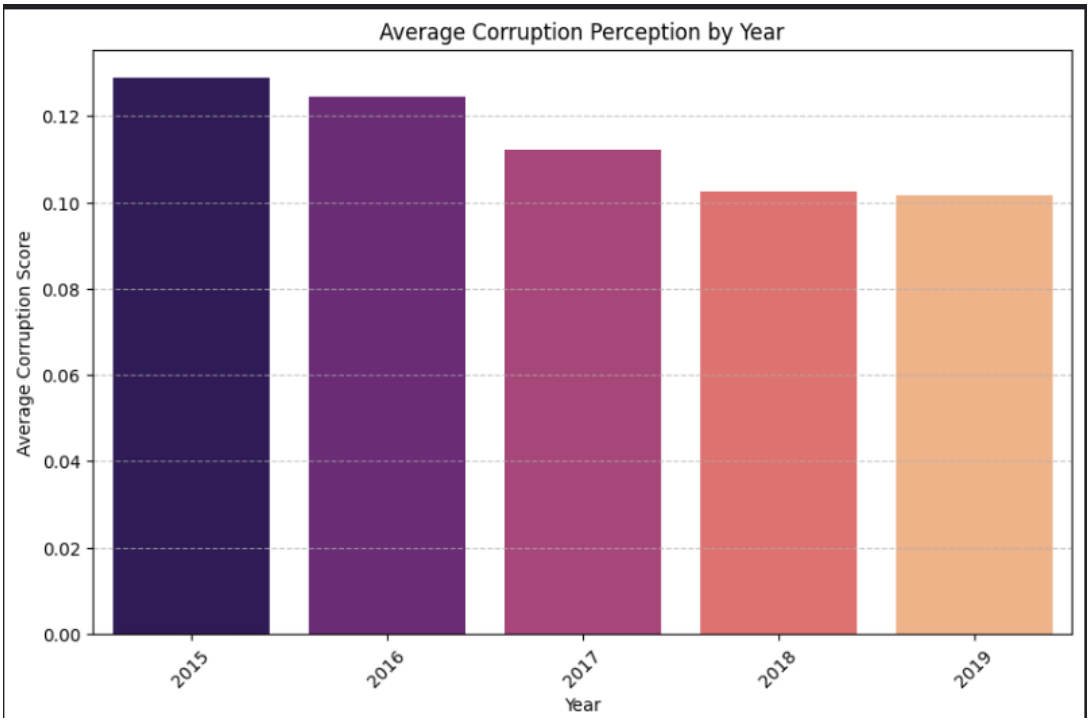

Figure 9: Average corruption perception scores from 2015 to 2019.

Description: Corruption perception has gradually declined over the years, reflecting possible improvements in governance and transparency.

4 Model Training and Evaluation

4.1 Data Splitting Strategy for Model Training

Before training regression models, the dataset was split into training and testing subsets to ensure robust evaluation and prevent overfitting.

4.1.1 Purpose of Data Splitting

- **Training Set (80%)**: Used for learning patterns in the data.
- **Testing Set (20%)**: Used for evaluating model performance on unseen data.
- This split ensures that the model generalizes well and is not overly fitted to the training dataset.

4.1.2 Splitting Approach

Two versions of the dataset were utilized:

- **Standardized Data**: Features were transformed using Z-score normalization.
- **Min-Max Scaled Data**: Features were normalized within a range of [0,1].

The dataset was split using an 80-20 ratio:

```
Standardized Data - Train Shape: (625, 5) Test Shape: (157, 5)
Min-Max Scaled Data - Train Shape: (625, 5) Test Shape: (157, 5)
```

Benefits of Data Splitting:

- Helps assess model performance on unseen data.
- Reduces overfitting by validating the model on separate test data.
- Provides a fair comparison among different models by ensuring consistency in evaluation.

4.2 Regression Models

The following regression models were trained and evaluated:

- **Linear Regression**
- **Ridge and Lasso Regression**
- **Random Forest Regressor**
- **XGBoost and Gradient Boosting**
- **LightGBM**
- **Support Vector Regression (SVR)**
- **Extra Trees and Bagging Regressors**
- **Bayesian Ridge**

4.3 Performance Metrics

Models were evaluated based on:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R-squared Score (R^2)

5 Model Training and Evaluation

5.1 Overview of Trained Models

To predict happiness scores, multiple regression models were trained using both standardized and min-max scaled data. The models included:

- **Linear Regression** - A simple baseline model assuming a linear relationship between features and happiness score.
- **Ridge and Lasso Regression** - Regularized models to prevent overfitting.
- **Random Forest Regressor** - An ensemble-based decision tree model for capturing complex interactions.
- **XGBoost and Gradient Boosting** - Advanced boosting techniques for improving prediction accuracy.
- **Support Vector Regression (SVR)** - A robust regression approach to minimize error.
- **Extra Trees and Bagging Regressors** - Ensemble techniques improving prediction variance.
- **Bayesian Ridge** - A probabilistic model incorporating prior knowledge.

5.2 Model Training and Hyperparameter Optimization

- Each model was trained on the standardized and min-max scaled datasets.
- Hyperparameter tuning was performed using Optuna to optimize model performance.
- K-fold cross-validation was applied to ensure robust model evaluation.

5.3 Performance Metrics

The models were evaluated based on the following metrics:

- **Mean Absolute Error (MAE)** - Measures the average absolute difference between actual and predicted values.
- **Mean Squared Error (MSE)** - Penalizes larger errors by squaring the differences.
- **Root Mean Squared Error (RMSE)** - Interpretable measure of prediction error.
- **R-squared Score** (R^2) - Indicates the proportion of variance explained by the model.

5.4 Model Performance Comparison

The table below summarizes the performance of all trained models:

Model	**MAE**	**MSE**	**RMSE**	R^2 **Score**
Linear Regression	0.4521	0.3432	0.5858	0.7178
Ridge Regression	0.4526	0.3441	0.5866	0.7170
Lasso Regression	0.4521	0.3432	0.5858	0.7178
Random Forest	0.4005	0.2707	0.5203	0.7774
XGBoost	0.4214	0.2758	0.5251	0.7733
Gradient Boosting	0.4133	0.2731	0.5226	0.7754
LightGBM	0.4380	0.2935	0.5417	0.7587
SVR	0.3956	0.2510	0.5010	0.7937
Extra Trees	0.3914	0.2502	0.5002	0.7943
Bagging Regressor	0.4011	0.2696	0.5192	0.7783
Bayesian Ridge	0.4524	0.3437	0.5862	0.7175

Table 2: Performance Comparison of Regression Models

Observations:

- **Extra Trees and SVR** achieved the best overall performance with the lowest RMSE and highest R^2 scores.
- **Random Forest and Bagging Regressor** performed well, demonstrating strong generalization.
- **Linear, Ridge, and Lasso Regression** showed similar performance but were outperformed by more advanced models.

5.5 Model Selection

Based on the evaluation, the Extra Trees Regressor and Support Vector Regression (SVR) were selected as the best models due to their superior performance in predicting happiness scores. These models achieved the lowest Root Mean Squared Error (RMSE) and the highest R^2 scores, indicating their strong predictive power.

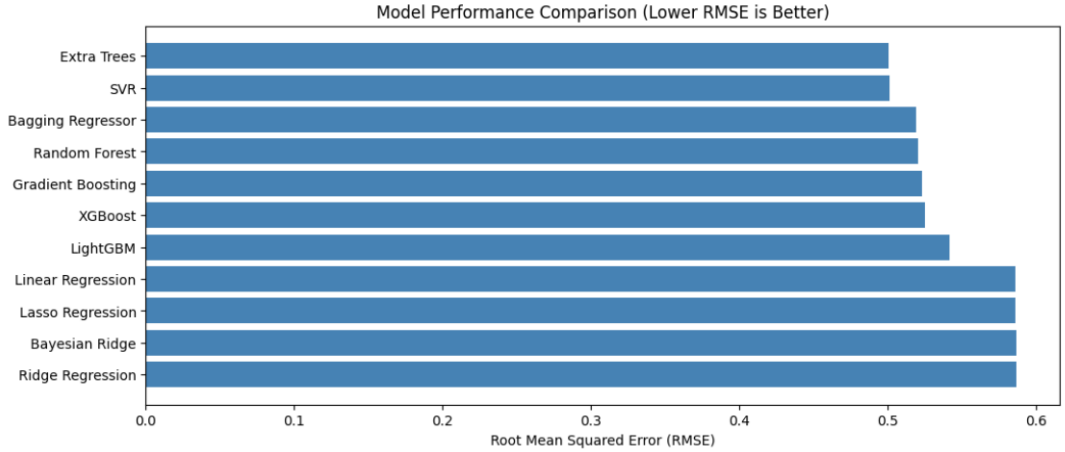

Figure 10: Model Performance Comparison (Lower RMSE is Better)

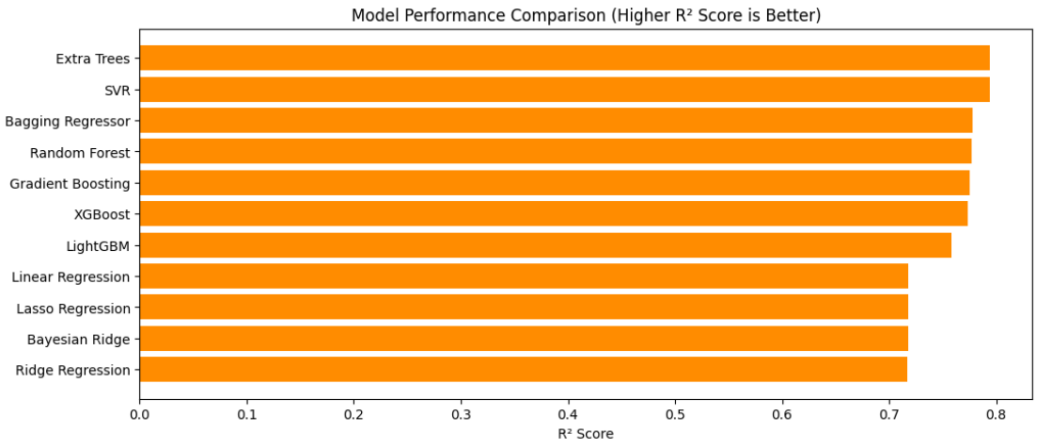

Figure 11: Model Performance Comparison (Higher R^2 Score is Better)

Discussion: - **Extra Trees Regressor** exhibited the lowest RMSE, making it highly effective in reducing prediction errors. - **SVR** demonstrated a strong balance between bias and variance, achieving a high R^2 score, indicating good generalization. - Both models outperformed traditional regression techniques such as Linear Regression, Ridge, and Lasso, which had higher RMSE values. - Tree-based models (Random Forest, Gradient Boosting, and XGBoost) also performed well, but Extra Trees proved to be the most effective. - Bayesian Ridge and Linear models had relatively lower performance due to their inability to capture complex relationships in the data.

These findings suggest that non-linear models, especially ensemble-based methods like Extra Trees and SVR, are more suitable for predicting happiness scores accurately.

6 Model Interpretation and Explainability

6.1 Feature Importance

Feature importance analysis using Extra Trees Regressor confirmed:

- **GDP per Capita** and **Life Expectancy** as the top predictors.
- **Freedom** and **Social Support** having significant impact.

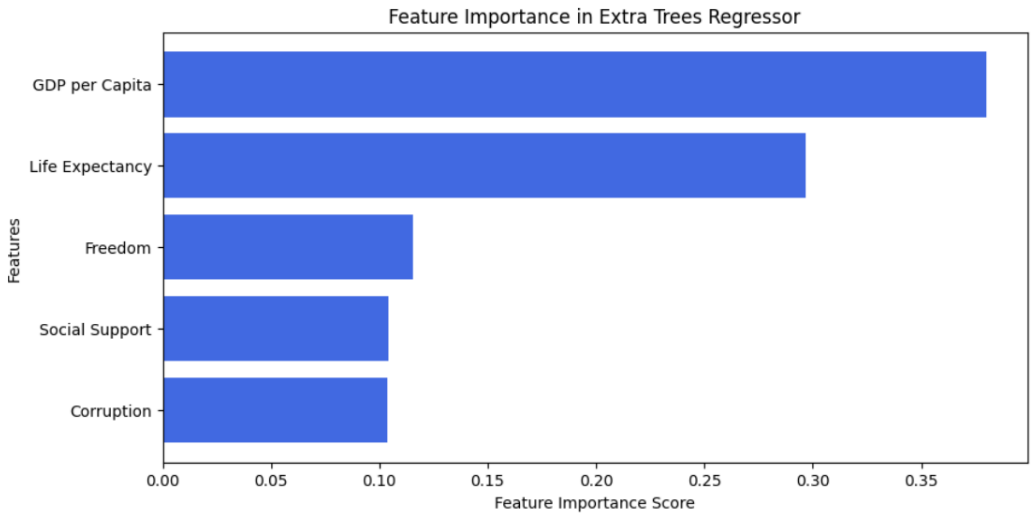

Figure 12: Feature Importance in Extra Trees Regressor

Discussion: The feature importance plot indicates that:

- **GDP per Capita** has the strongest influence on happiness scores, reaffirming the economic aspect of well-being.
- **Life Expectancy** significantly contributes to happiness, emphasizing the importance of healthcare and longevity.
- **Freedom** and **Social Support** play substantial roles, demonstrating that societal and political factors influence happiness levels.
- **Corruption** has a lower but notable impact, highlighting governance quality as a factor in perceived well-being.

6.2 SHAP Analysis

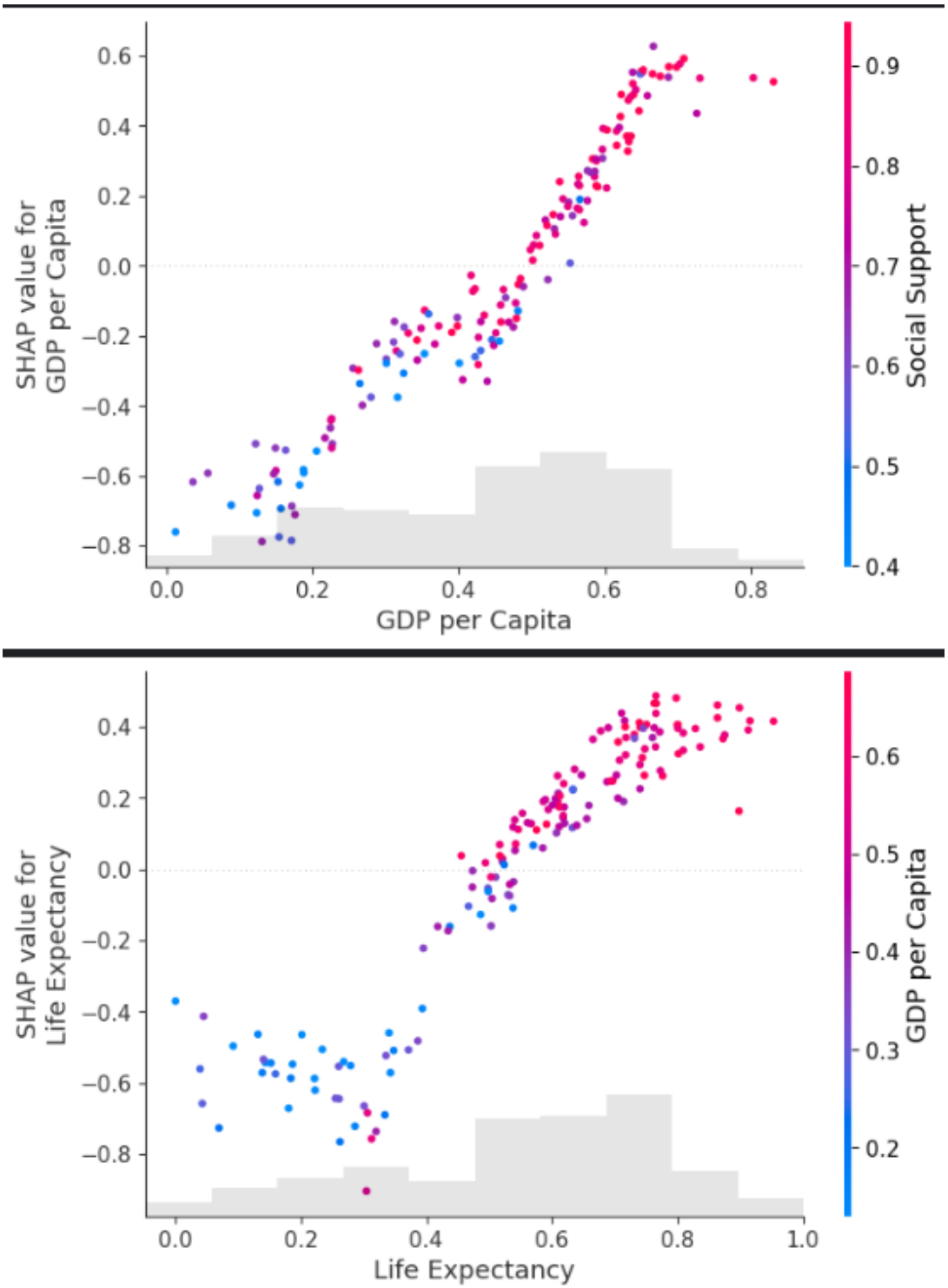

Figure 13: SHAP Scatter Plots: The impact of GDP per Capita and Life Expectancy on predictions. Points are color-coded to indicate the values of other influencing features. Higher GDP per Capita and Life Expectancy strongly contribute to higher happiness scores.

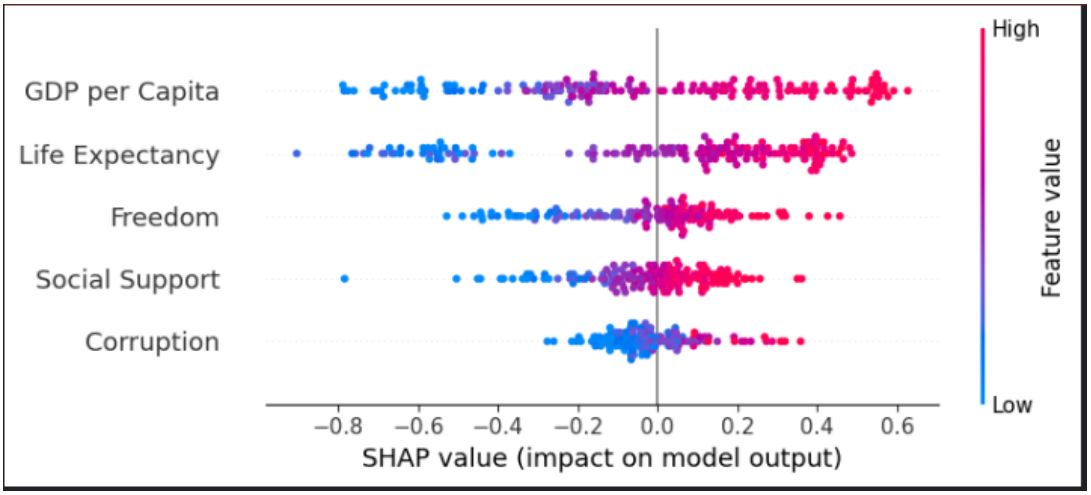

Figure 14: SHAP Summary Plot: The overall feature importance for predictions. Features with higher SHAP values (such as GDP per Capita and Life Expectancy) have the most significant influence on the model's predictions.

- SHAP summary plots revealed global feature contributions, highlighting the most influential features.
- Individual SHAP force plots explained specific model predictions, helping interpret how each factor contributes to a happiness score.
- The scatter plots illustrate the positive correlation between GDP per Capita, Life Expectancy, and Happiness Score.

7 Conclusion and Future Work

This study aimed to analyze the key factors influencing happiness and develop predictive models to estimate happiness scores based on socio-economic indicators. The exploratory data analysis highlighted significant relationships between happiness and variables such as GDP per Capita, Social Support, Life Expectancy, and Freedom. Feature selection further confirmed these variables as the most influential factors.

Through rigorous model evaluation, Support Vector Regression (SVR) and Extra Trees Regressor were identified as the best-performing models, achieving the lowest RMSE and highest R-squared scores. The performance analysis indicated that ensemble-based models provided better generalization due to their ability to capture complex relationships in the data.

SHAP analysis provided interpretability by illustrating how individual features impacted predictions. The insights gained from SHAP visualizations confirmed that higher GDP per Capita and longer Life Expectancy positively influenced happiness scores, reinforcing findings from previous studies on global well-being.

Future work can expand on this study by incorporating additional socio-economic and psychological factors that might influence happiness, such as employment rates, mental health indicators, and political stability. Additionally, experimenting with deep learning architectures such as neural networks could provide improved predictive performance. Further research can also explore causal relationships through time-series forecasting and policy simulations to assess how changes in key factors might affect happiness trends globally.