DSA 210 Project Final Report

Can SDG Performance Predict Happiness? A Machine Learning-Based Classification Study

Safiye Nur Narman

32411

INTRODUCTION

The Sustainable Development Goals (SDGs) have become a universal framework for evaluating a country's social, economic, and environmental progress. At the same time, happiness indices, such as the World Happiness Score, offer insights into the well-being and life satisfaction of populations. Understanding the relationship between sustainable development and happiness can offer valuable perspectives for policymakers and researchers alike.

This study investigates whether countries with lower SDG scores tend to have lower happiness scores. In particular, it frames the problem as a binary classification task: distinguishing between countries with high and low SDG performance using features such as their Happiness Score and the corresponding year. By applying multiple machine learning models—including logistic regression, random forest, and XGBoost—the study aims to explore whether happiness can serve as a meaningful indicator of a country’s sustainable development performance.

The goal is to evaluate the predictive capacity of happiness metrics and temporal data in identifying underperforming countries in terms of SDG achievement.

DATASET OVERVIEW

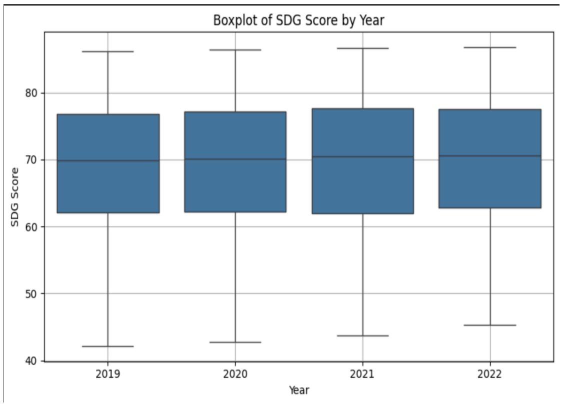
The dataset used in this study combines information on countries’ Sustainable Development Goal (SDG) scores, Happiness Scores, and the corresponding years. The SDG scores provide a composite index measuring progress toward the United Nations’ 17 development goals, while the Happiness Score reflects national well-being based on factors such as income, social support, and life expectancy.

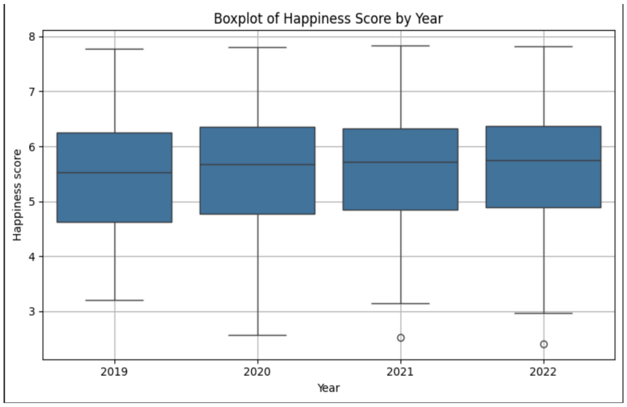
After necessary cleanings, each row in the dataset represents a country-year observation, allowing for a temporal dimension to be included in the analysis. The two key features selected for modeling were the **Happiness Score** and the **Year**, while the **SDG Score** served as the basis for the binary classification labels.

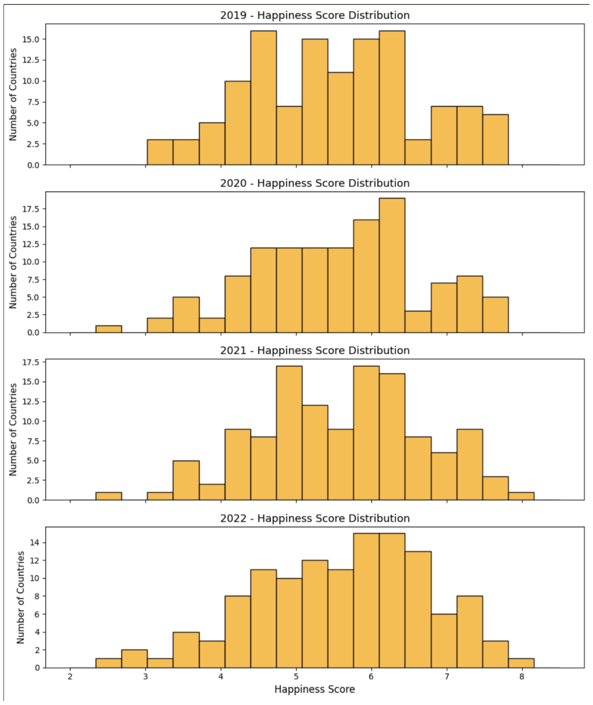
EXPLORATORY DATA ANALYSIS (EDA) & VISUALIZATION

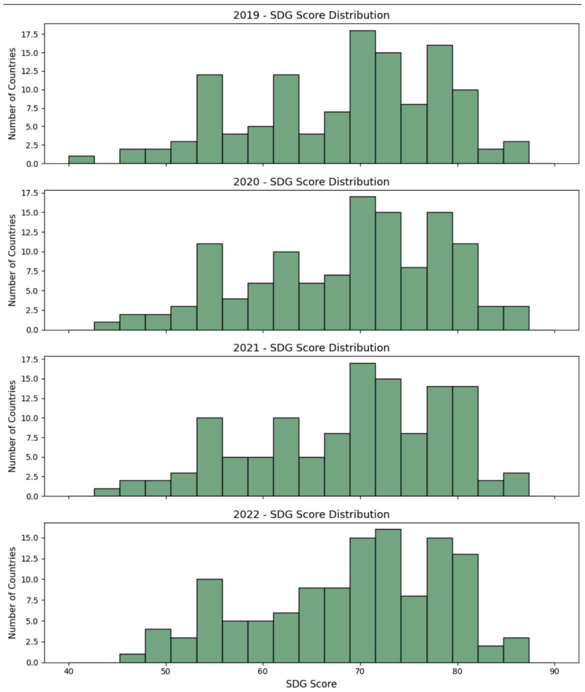
3.1 Distributions by Year

Histograms and boxplots were employed to visualize the annual distributions of SDG Scores and Happiness Scores. SDG Scores remained relatively stable, showing gradual improvements, while Happiness Scores exhibited fluctuations, reaching a peak in 2021 followed by a slight decline in 2022, possibly influenced by global events such as the COVID-19 pandemic or economic instability.



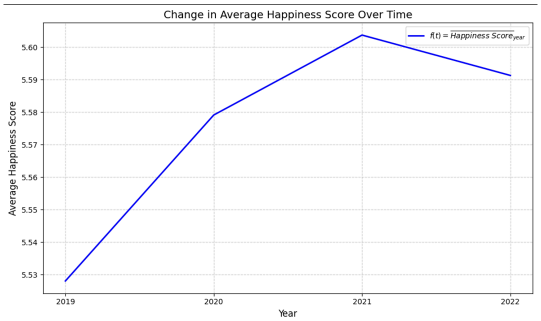


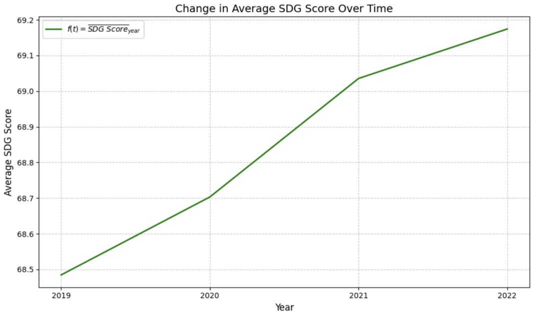




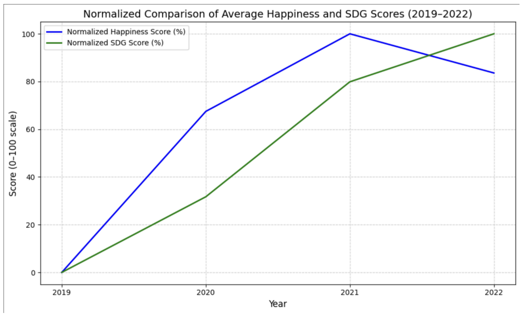
3.2 Trends in Average Scores

Line plots were utilized to illustrate average score trends over the years. The SDG Score steadily improved between 2019 and 2022, whereas the Happiness Score increased until 2021 and then slightly declined. This discrepancy suggests sustainable development progress does not uniformly translate to enhanced perceived happiness.



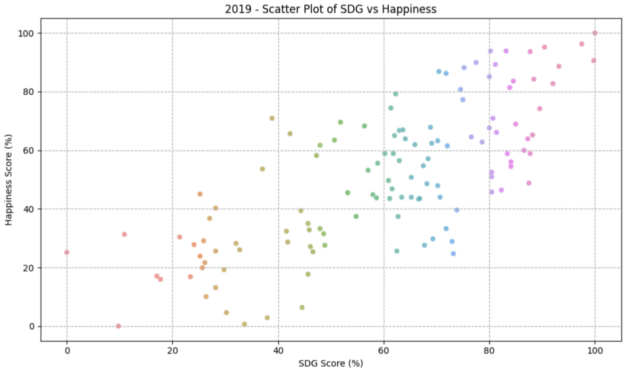


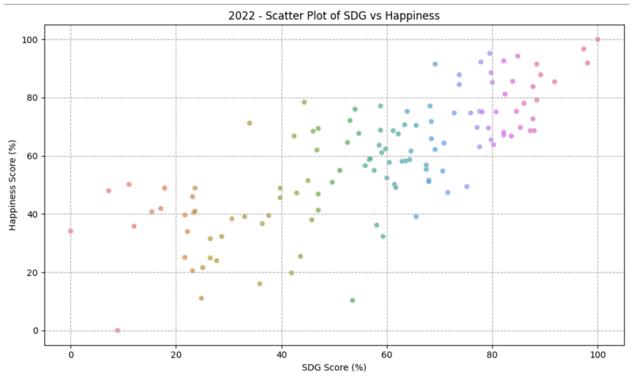
Also conducted normalized analysis which revealed a consistent and linear improvement in SDG Scores, whereas improvements in Happiness Scores plateaued and subsequently decreased, highlighting potential divergences between objective development metrics and subjective well-being.



3.3 Relationship Between SDG and Happiness

Scatter plots indicated a positive relationship between SDG Scores and Happiness Scores, with higher SDG-performing countries generally reporting higher levels of happiness. While not strictly linear, the data points collectively indicated a general upward correlation.





3.4 Implications for Modeling

Insights from the EDA informed the decision to frame the problem as a binary classification task, distinguishing countries based on their SDG Score relative to the median. The identified relationship between SDG and Happiness Scores supported the use of Happiness Score and Year as key predictive features for subsequent machine learning modeling.

HYPOTHESIS TESTING

4.1 Correlation Analysis

- **Null Hypothesis (H_0):** There is no significant correlation between SDG Scores and Happiness Scores.

- **Alternative Hypothesis (H_1):** There is a significant correlation between SDG Scores and Happiness Scores.

To evaluate the relationship between SDG Scores and Happiness Scores, the following hypotheses were tested:

The Pearson correlation analysis consistently demonstrated a strong and statistically significant positive relationship ($p < 0.05$) from 2019 to 2022. This result supports the idea that countries achieving better sustainable development performance tend to experience higher levels of happiness.

4.2 Difference in Means Analysis

To test the differences in happiness scores between high and low SDG performers, the following hypotheses were used:

- **Null Hypothesis (H_0):** There is no significant difference in happiness scores between high and low SDG performers.

- **Alternative Hypothesis (H_1):** There is a significant difference in happiness scores between high and low SDG performers.

Welch's T-tests showed significant differences ($p < 0.05$) in happiness scores between the two groups across all years. This finding indicates that higher SDG scores are reliably associated with increased national happiness.

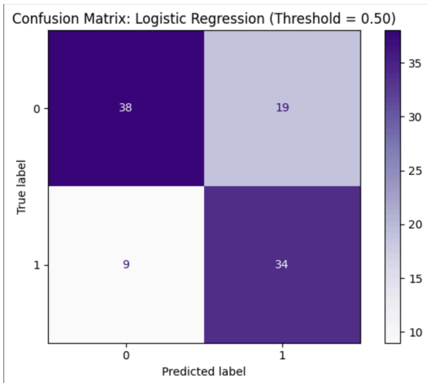
4.3 Summary

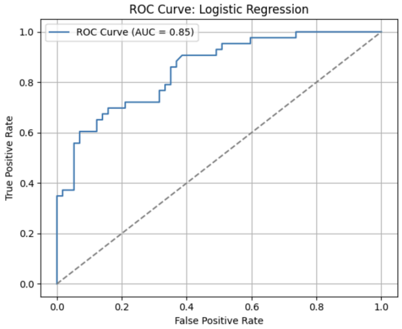
The hypothesis tests collectively affirm that better sustainable development outcomes are positively correlated with higher happiness levels, reinforcing the importance of SDG initiatives for enhancing national well-being.

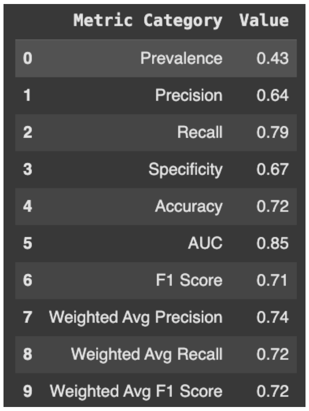
MACHINE LEARNING

5.1 Logistic Regression

The Logistic Regression model provided the strongest predictive performance among the tested methods. It effectively distinguished between countries with high and low SDG scores, demonstrating robust accuracy and strong discriminatory capability. The model was particularly effective at correctly identifying lower-performing countries, crucial for targeted policy interventions, despite occasionally misclassifying higher performers.



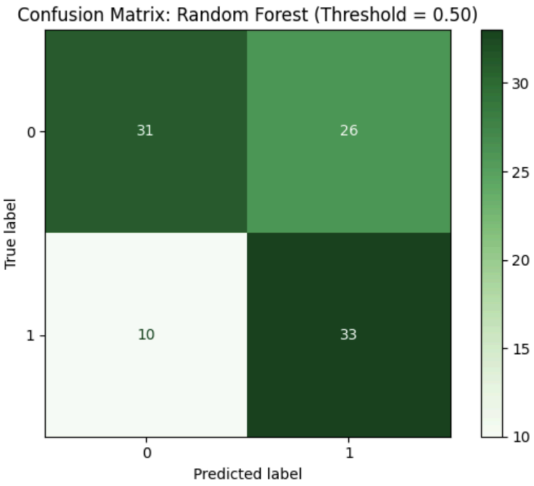


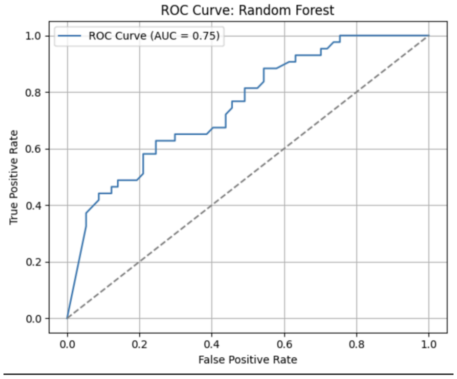


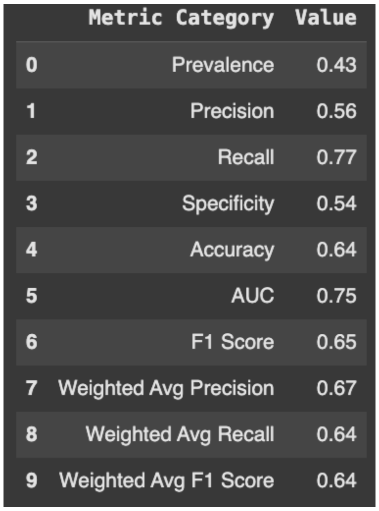
	Metric Category	Value
0	Prevalence	0.43
1	Precision	0.64
2	Recall	0.79
3	Specificity	0.67
4	Accuracy	0.72
5	AUC	0.85
6	F1 Score	0.71
7	Weighted Avg Precision	0.74
8	Weighted Avg Recall	0.72
9	Weighted Avg F1 Score	0.72

5.2 Random Forest

The Random Forest model showed moderate effectiveness in classification. While effective at identifying lower-performing countries, it encountered difficulties distinguishing top performers, resulting in higher false-positive rates. This model provided a balanced but less precise alternative compared to Logistic Regression, suitable for broader analyses where absolute precision is less critical.



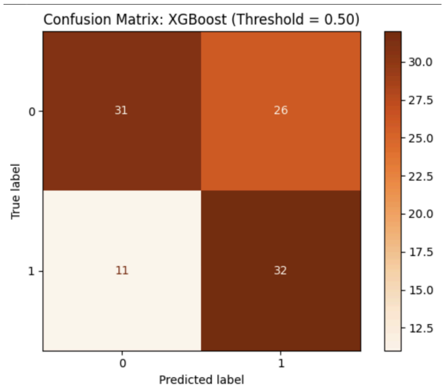


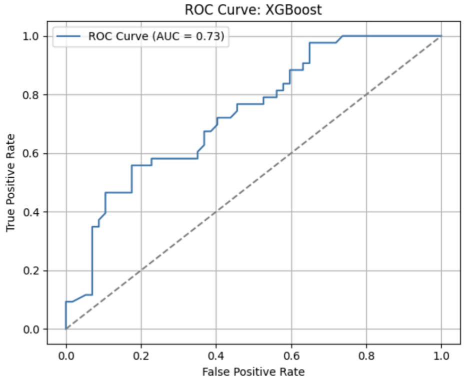


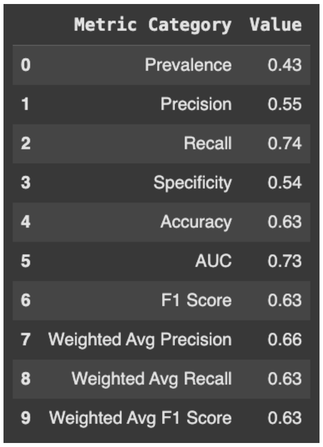
	Metric Category	Value
0	Prevalence	0.43
1	Precision	0.56
2	Recall	0.77
3	Specificity	0.54
4	Accuracy	0.64
5	AUC	0.75
6	F1 Score	0.65
7	Weighted Avg Precision	0.67
8	Weighted Avg Recall	0.64
9	Weighted Avg F1 Score	0.64

5.3 XGBoost

The XGBoost model offered moderate performance, displaying good generalization across both classes but exhibiting a slightly higher rate of misclassification compared to Logistic Regression. Its performance was reliable but somewhat limited in precision, making it useful in scenarios tolerant of moderate uncertainty.







	Metric Category	Value
0	Prevalence	0.43
1	Precision	0.55
2	Recall	0.74
3	Specificity	0.54
4	Accuracy	0.63
5	AUC	0.73
6	F1 Score	0.63
7	Weighted Avg Precision	0.66
8	Weighted Avg Recall	0.63
9	Weighted Avg F1 Score	0.63

5.4 Summary

Among the three models tested, Logistic Regression demonstrated superior predictive power and overall balance. Random Forest and XGBoost models performed moderately well but presented some limitations regarding precision. Logistic Regression is recommended for practical applications where correctly identifying lower-performing countries is a priority.

CONCLUSION

This study integrated exploratory data analysis, hypothesis testing, and machine learning to comprehensively assess the relationship between Sustainable Development Goals (SDGs) performance and national happiness levels. The combined findings strongly affirm that better sustainable development outcomes are positively correlated with higher levels of national happiness.

The exploratory analysis provided initial insights, clearly demonstrating trends and relationships over multiple years. Subsequent hypothesis testing statistically validated these observations, confirming that the association between SDG performance and happiness is both strong and significant.

Machine learning models, particularly Logistic Regression, effectively classified countries based on SDG performance using happiness scores and year as predictive indicators. The robust performance of these models underscores happiness as a valuable and meaningful indicator of sustainable development success.

Key Insights:

- Strong positive correlation exists between sustainable development (SDG scores) and happiness.
- Higher SDG performance consistently aligns with higher happiness levels across different countries.
- Machine learning, especially logistic regression, provides effective tools for identifying countries needing targeted improvement.

Social Implications:

These findings highlight the critical importance of investing in sustainable development initiatives as a pathway to enhancing overall national well-being. Policymakers and stakeholders are encouraged to leverage these insights to guide sustainable and inclusive development strategies that meaningfully improve citizens' lives.

Recommendations for Future Work:

- Further studies could include additional socio-economic variables, such as education, healthcare quality, and economic inequality, to enhance model predictive power.
- Broader application of these findings can inspire international cooperation, emphasizing sustainable development as essential for global well-being and stability.