M10.6 Group 5 Final Project Report
Predicting Happiness Score: The Role of Socio-Cultural and Political Factors in Regional Well-being

Almy, Britton
University of Tennessee

Gupta, Atul
University of Tennessee

Proctor, Mary
University of Tennessee

Rangel, Jarrod
University of Tennessee

Abstract —This project uses machine learning models to predict national happiness scores based on socio-cultural, economic, and political factors. We applied and compared Linear Regression, SVR, Random Forest, XGBoost, and Deep Learning techniques. Ensemble models—particularly XGBoost—demonstrated the highest accuracy. By incorporating expanded features like internet access and mental health, our approach offers a more comprehensive view of global well-being. The results support the value of machine learning in informing happiness-related policymaking.

Introduction

The World Happiness Report is an influential global survey that assesses the state of happiness across numerous countries. First published in 2012, the report has consistently provided critical insights into how happiness indicators can guide policymaking worldwide. Utilizing data primarily from the Gallup World Poll, the report's rankings are derived from the Cantril ladder method, where respondents rate their current lives on a scale from 0 (worst possible life) to 10 (best possible life). Factors influencing these happiness scores include economic production, social support, life expectancy, freedom, absence of corruption, and generosity, with each contributing uniquely to national happiness.

The concept of "Dystopia," a hypothetical country with the lowest values for these six key factors, provides a baseline to positively compare other countries, highlighting variations in happiness. Additionally, residuals or unexplained components indicate the extent to which these six factors do not fully explain happiness scores, adding depth to the analysis.

This project leverages the World Happiness dataset to develop predictive machine learning models, aiming to identify the socio-cultural and political factors that most significantly affect happiness scores.

Data Sourcing and Preprocessing

The World Happiness Report dataset was meticulously preprocessed. Missing values were identified, and appropriately managed, categorical variables were one-hot encoded, and numerical features were standardized. The dataset was then partitioned into training and testing sets to facilitate effective model evaluation.

The preprocessing steps involved several critical stages:

- Data Cleaning: Missing values were identified and handled through appropriate imputation techniques or removal, ensuring data integrity. Duplicate records were eliminated.
- Feature Engineering: Column names were standardized, and relevant derived metrics were created to capture deeper insights.
- Data Transformation: Categorical variables were transformed via one-hot encoding, and numerical features were standardized to maintain consistency and comparability.
- Exploratory Data Analysis (EDA): Statistical summaries and visualizations, such as correlation heatmaps, histograms, and scatter plots, were generated to reveal underlying patterns, relationships, and outliers in the data.

These preprocessing steps were fundamental in preparing the dataset for accurate, meaningful predictive analysis.

Table 1 Raw Data Example

	0	1	2	3	4
Country	China	UK	Brazil	France	China
Year	2022	2015	2009	2019	2022
Happiness_Score	4.39	5.49	4.65	5.2	7.28
GDP_per_Capita	44984.68	30814.59	39214.84	30655.75	30016.87
Social_Support	0.53	0.93	0.03	0.77	0.05
Healthy_Life_Expectancy	71.11	63.14	62.36	78.94	50.33
Freedom	0.41	0.89	0.01	0.98	0.62
Generosity	-0.05	0.04	0.16	0.25	0.18
Corruption_Perception	0.83	0.84	0.59	0.63	0.92
Unemployment_Rate	14.98	19.46	16.68	2.64	7.7
Education_Index	0.52	0.83	0.95	0.7	0.92
Population	1311940760	1194240877	731100898	1293957314	1432971455
Urbanization_Rate	78.71	50.87	48.75	81.78	82.39
Life_Satisfaction	8.88	5.03	5.22	5.69	6.33
Public_Trust	0.34	0.72	0.23	0.68	0.5
Mental_Health_Index	76.44	53.38	82.4	46.87	60.38
Income_Inequality	46.06	46.43	31.03	57.65	28.54
Public_Health_Expenditure	8.92	4.43	3.78	4.43	7.66
Climate_Index	62.75	53.11	33.3	90.59	59.33
Work_Life_Balance	8.59	8.76	6.06	6.36	3.0
Internet_Access	74.4	91.74	71.8	86.16	71.1
Crime_Rate	70.3	73.32	28.99	45.76	65.67
Political_Stability	0.29	0.76	0.94	0.48	0.12
Employment_Rate	61.38	80.18	72.65	55.14	51.55

Note, table data has been rotated for visibility with Column on the left and rows proceeding left to rate labeled 0-4

Methods

The selected regression algorithms included Support Vector Regression (SVR) and Deep Learning Scalar Regression. Both models were trained and fine-tuned to optimize predictive accuracy. Model performance was evaluated using a holdout test set, with metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared serving as evaluation criteria.

Each team member also tested random subsets of the dataset independently, applying both regression models. The individual results were later compiled and compared to assess consistency across subsets and identify any discrepancies or unique insights.

- Correlation analysis to inform feature selection.
- Training and validation splits for individual subsets.
- Application of Linear Regression, SVR, Random Forest, XGBoost, and Deep Learning models.
- Normalization and scaling of data to ensure model accuracy and comparability.

The following timeline outlines our structured approach for systematically acquiring, analyzing, and modeling the dataset over a four-week period. Each phase includes key tasks such as data preprocessing, exploratory analysis, individual testing, and rigorous validation to ensure thorough evaluation. By clearly defining roles and objectives each week, we aim to optimize our model development and deliver robust, validated results.

Week 1:

- Dataset acquisition and initial preprocessing.
- Exploratory Data Analysis (EDA) to understand dataset characteristics.
- Subset definitions and assignments for individual team member analysis.

Week 2:

- Detailed correlation analysis to guide feature selection, identifying and eliminating highly correlated or irrelevant variables.
- Initial development and training of regression models, including Linear Regression, Support Vector Regression (SVR), Random Forest, XGBoost, and Deep Learning Scalar Regression.
- Initiation of independent testing by each team member on their assigned dataset subsets.

Week 3:

- Model parameter fine-tuning and validation checks to optimize predictive performance.
- Completion of individual subset testing and documentation of findings by each team member.

Week 4:

- Consolidation and analysis of individual testing results.
- Evaluation of model performance using defined metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R²).
- Compilation of comprehensive findings into the final report, highlighting actionable insights and identifying model limitations for future improvements.

Data Cleanup

During the data cleanup process, rows containing missing values were removed entirely to ensure the integrity of the dataset. Additionally, columns with missing numeric values were treated by replacing these gaps with the mean of the available data, preserving the consistency and minimizing bias. These steps helped enhance the dataset's reliability for subsequent analysis.

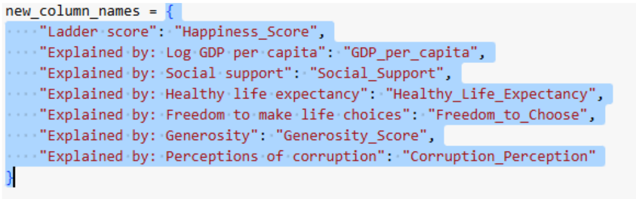
Exploratory Data Analysis (EDA)

To better understand the underlying structure and characteristics of the dataset, an Exploratory Data Analysis (EDA) was conducted. This process involved inspecting the data for completeness, renaming columns for clarity, analyzing the distribution of key features, and examining relationships between variables through correlation analysis. The EDA helped to identify important patterns, highlight potential anomalies, and inform the approach for subsequent modeling and interpretation.

Rename of Columns

The initial step in the data preprocessing involved renaming columns to ensure clarity and consistency throughout the dataset. This was essential to facilitate accurate interpretation and analysis of the data.

Fig. 1. Column Code



```
new_column_names = {
    "Ladder score": "Happiness_Score",
    "Explained by: Log GDP per capita": "GDP_per_capita",
    "Explained by: Social support": "Social_Support",
    "Explained by: Healthy life expectancy": "Healthy_Life_Expectancy",
    "Explained by: Freedom to make life choices": "Freedom_to_Choose",
    "Explained by: Generosity": "Generosity_Score",
    "Explained by: Perceptions of corruption": "Corruption_Perception"
}
```

Correlation of Features

The correlation between various features was examined, with a particular focus on their relationship to the Happiness_Score. Findings include:

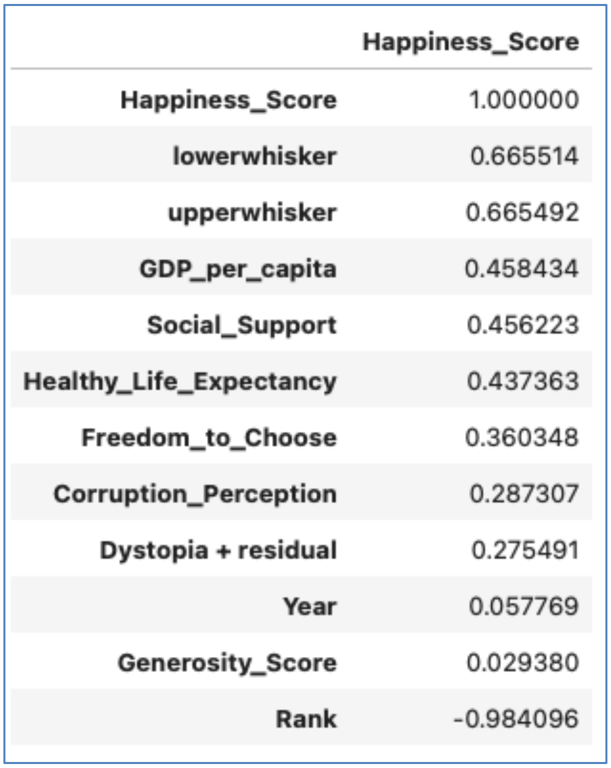
The Lower Whisker and Upper Whisker demonstrated a notable positive correlation with

Happiness_Score, with a correlation coefficient of approximately 0.665. This suggests that the distribution of the data, as represented by these whiskers, plays a significant role in influencing happiness scores.

Other factors such as GDP_per_capita, Social_Support, and Healthy_Life_Expectancy show moderate positive correlations with Happiness_Score (ranging from 0.45 to 0.46), this indicates that higher levels of economic prosperity, social support, and health expectancy are associated with higher happiness scores.

A highly significant negative correlation was observed between Rank and Happiness_Score, with a coefficient of -0.984. This negative relationship suggests that as countries rank higher (i.e., with better happiness scores), their happiness index values tend to increase, reflecting an inverse relationship with the rankings.

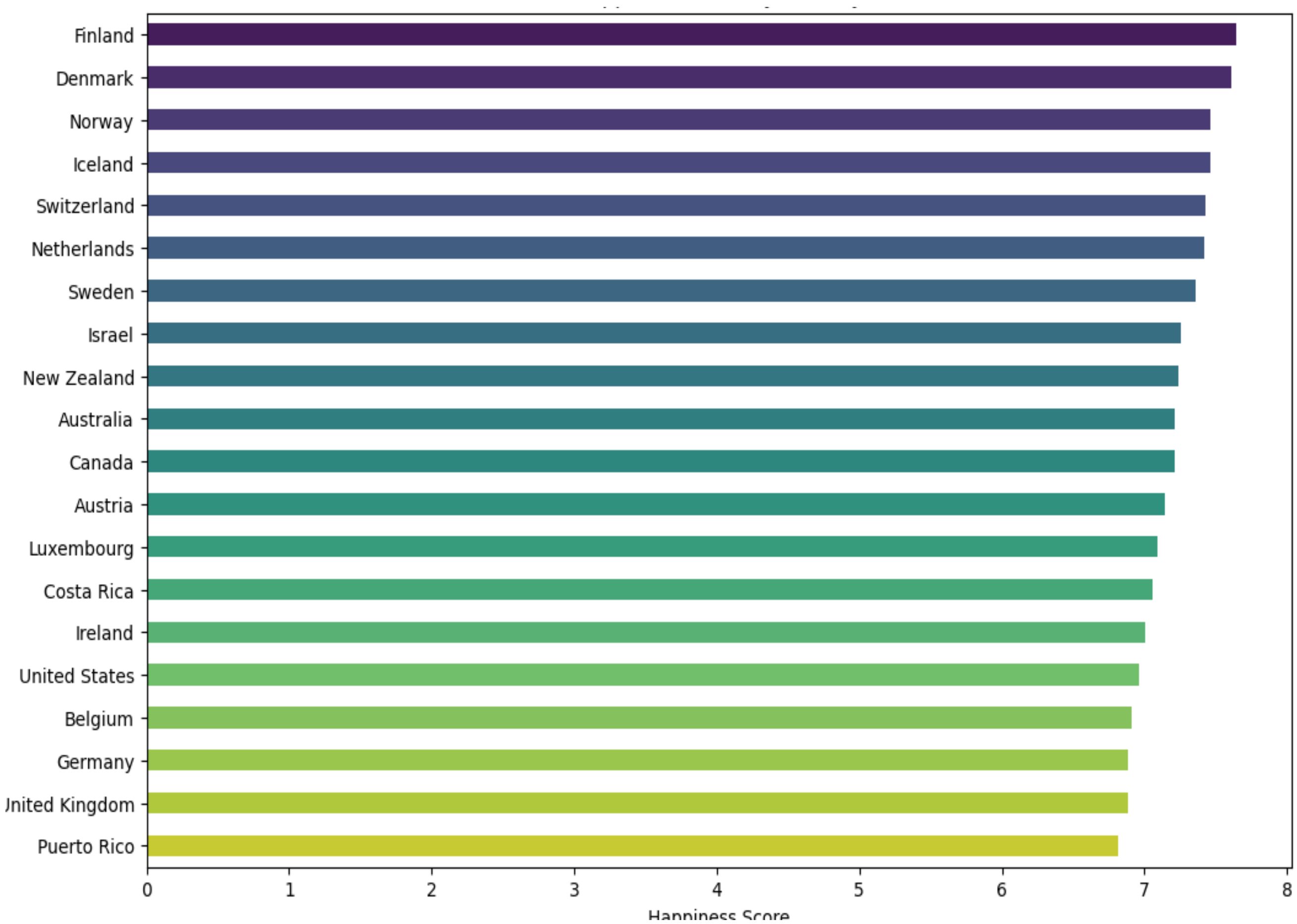
Table 2 Scores



	Happiness_Score
Happiness_Score	1.000000
lowerwhisker	0.665514
upperwhisker	0.665492
GDP_per_capita	0.458434
Social_Support	0.456223
Healthy_Life_Expectancy	0.437363
Freedom_to_Choose	0.360348
Corruption_Perception	0.287307
Dystopia + residual	0.275491
Year	0.057769
Generosity_Score	0.029380
Rank	-0.984096

Analysis of Data Distribution

The analysis of the data reveals that happiness scores are generally right-skewed, with most countries scoring between 6 and 7, and a few high-scoring nations in northern Europe. GDP per capita shows a moderate positive correlation with happiness, though it's not the only determinant. Social support and healthy life expectancy have a more even distribution, while perceptions of freedom vary. Generosity has little impact on happiness, while lower corruption levels correlate with higher happiness. Overall, wealth, social support, and health are key factors, but freedom and corruption also play significant roles.Happiness Score by Country



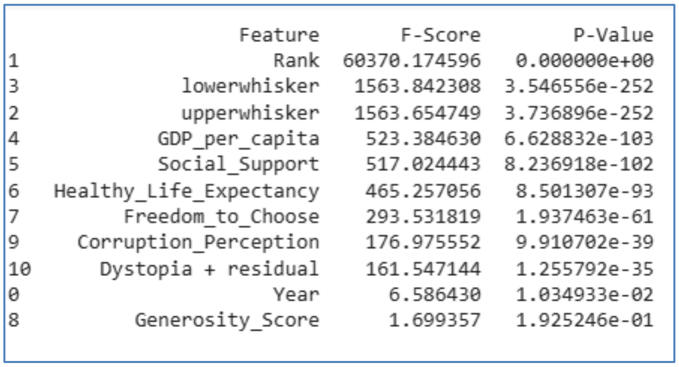
Feature Scaling

Since machine learning models like Support Vector Regression (SVR) and other algorithms can be sensitive to the scale of data, StandardScaler is applied to scale the features to have a mean of 0 and a standard deviation of 1. This helps in ensuring that the model does not overfit to certain features due to scale differences.

f_regression

f_regression is a statistical test used in the context of feature selection, specifically for regression tasks. It is part of the process where the goal is to identify the most relevant features for predicting the target variable.

Fig. 2. Feature Selection



```
                   Feature        F-Score        P-Value
1                     Rank   60370.174596   0.000000e+00
3            lowerwhisker     1563.842308  3.546556e-252
2            upperwhisker     1563.654749  3.736896e-252
4           GDP_per_capita     523.384630  6.628832e-103
5           Social_Support     517.024443  8.236918e-102
6  Healthy_Life_Expectancy     465.257056   8.501307e-93
7        Freedom_to_Choose     293.531819   1.937463e-61
9    Corruption_Perception     176.975552   9.910702e-39
10      Dystopia + residual    161.547144   1.255792e-35
0                     Year       6.586430   1.034933e-02
8         Generosity_Score       1.699357   1.925246e-01
```

Selection of Models

The following models were considered for this regression task:

Linear Regression

Why Considered: A simple and interpretable model that assumes a linear relationship between the

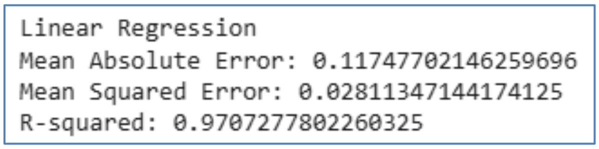
input features and the target variable.
Pros:

- Easy to implement and interpret.
- Fast to train and evaluate.

Cons:

- Assumes linearity between features and target variable, which might not always be the case.

Fig. 3. Linear regression Results



```
Linear Regression
Mean Absolute Error: 0.11747702146259696
Mean Squared Error: 0.02811347144174125
R-squared: 0.9707277802260325
```

Support Vector Regression (SVR)

Why Considered: SVR is well-suited for situations where there is no clear linear relationship between features and the target. It can also model non-linear relationships using kernel tricks.

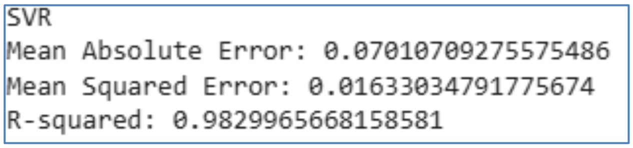
Pros:

- Effective for high-dimensional data and non-linear relationships.
- Robust to overfitting, especially in high-dimensional spaces.

Cons:

- Computationally expensive and sensitive to the choice of kernel
- Requires feature scaling (which we've done using StandardScaler)

Fig. 4. SVR Results



```
SVR
Mean Absolute Error: 0.07010709275575486
Mean Squared Error: 0.01633034791775674
R-squared: 0.9829965668158581
```

Random Forest Regression

Why Considered: A robust, ensemble method that can capture non-linear relationships. It is a tree-based method that combines multiple decision trees for more accurate predictions.

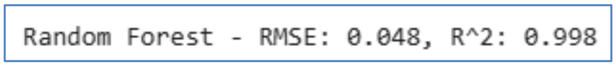
Pros:

- Can model complex, non-linear relationships.
- Less sensitive to outliers and overfitting compared to individual decision trees.

Cons:

- Requires more computational resources.
- Less interpretable compared to linear models.

Fig. 5. Random Forest Results



```
Random Forest - RMSE: 0.048, R^2: 0.998
```

Gradient Boosting Machines (GBM)

Why Considered: A powerful ensemble learning method that builds decision trees sequentially to improve predictive performance.

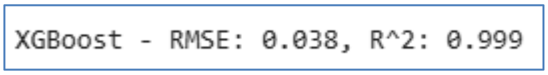
Pros:

- Strong predictive accuracy.
- Handles missing values and outliers well.

Cons:

- Computationally expensive.
- Requires careful tuning of hyperparameters.

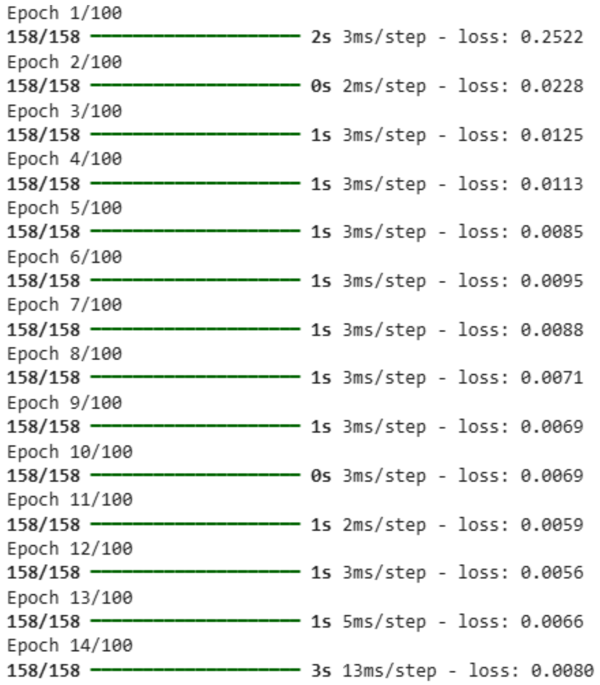
Fig. 6. GBM Results



```
XGBoost - RMSE: 0.038, R^2: 0.999
```

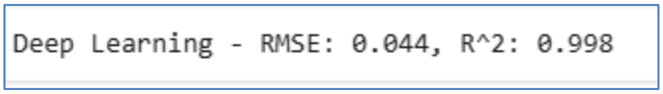
Deep Learning

Fig. 7. Epoch Progress



```
Epoch 1/100
158/158 ━━━━━━━━━━━━━━━━━━━━ 2s 3ms/step - loss: 0.2522
Epoch 2/100
158/158 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - loss: 0.0228
Epoch 3/100
158/158 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - loss: 0.0125
Epoch 4/100
158/158 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - loss: 0.0113
Epoch 5/100
158/158 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - loss: 0.0085
Epoch 6/100
158/158 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - loss: 0.0095
Epoch 7/100
158/158 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - loss: 0.0088
Epoch 8/100
158/158 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - loss: 0.0071
Epoch 9/100
158/158 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - loss: 0.0069
Epoch 10/100
158/158 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0069
Epoch 11/100
158/158 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - loss: 0.0059
Epoch 12/100
158/158 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - loss: 0.0056
Epoch 13/100
158/158 ━━━━━━━━━━━━━━━━━━━━ 1s 5ms/step - loss: 0.0066
Epoch 14/100
158/158 ━━━━━━━━━━━━━━━━━━━━ 3s 13ms/step - loss: 0.0080
```

Fig. 8. Deep Learning Results



```
Deep Learning - RMSE: 0.044, R^2: 0.998
```

Training & Testing

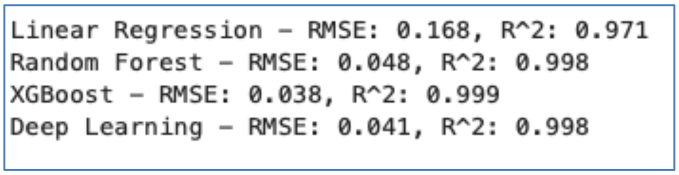
The dataset was partitioned into 80% for training and 20% for testing to ensure consistent evaluation across all models. Initially, the features (X) were separated from the target variable (y), with "Happiness_Score" chosen as the target. To prevent data leakage, feature scaling was performed. Additionally a fixed random state was used during the training to ensure reproducibility. This standardized approach allowed for a comparison analysis between the selected models.

Model performance is evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score, comparing all models performances.

Model Performance

From this comparison, XGBoost outperforms the other models, achieving the lowest RMSE (0.038) and the highest R^2 (0.999), indicating excellent predictive accuracy. The Random Forest and Deep Learning models also show strong performance with RMSE values of 0.048 and 0.041, respectively. The Linear Regression model has the highest RMSE and slightly lower R^2, suggesting it is less effective compared to the other models in this case.

Fig. 9. Model Results



```
Linear Regression - RMSE: 0.168, R^2: 0.971
Random Forest - RMSE: 0.048, R^2: 0.998
XGBoost - RMSE: 0.038, R^2: 0.999
Deep Learning - RMSE: 0.041, R^2: 0.998
```

Fig. 10. Actual vs Predicted Random Forest

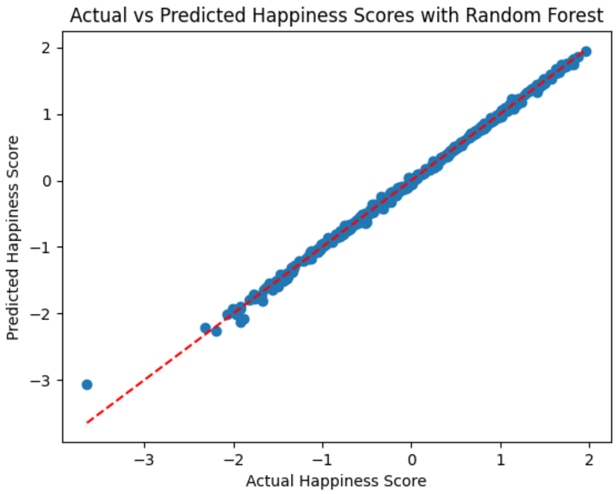
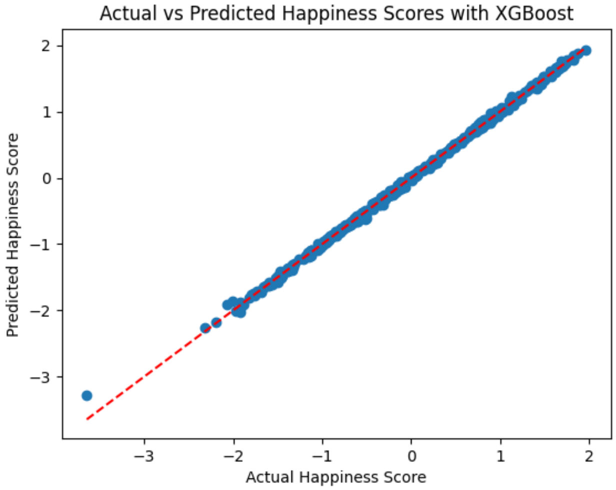


Fig. 11. Scatter Plot with XGBoost



Data Limitations

While this study provides meaningful insights into factors influencing national happiness, several data-related limitations must be acknowledged. First, the dataset relies heavily on self-reported survey responses, which can introduce subjective bias and inconsistencies across countries due to cultural differences in expressing well-being. Moreover, the sampling methodology of the Gallup World Poll may not uniformly represent all demographic groups within each country, potentially skewing national-level happiness scores.

Second, some variables that are likely relevant to happiness—such as access to mental health services, quality of education, or degree of income mobility—are either missing or poorly captured in the dataset. The inclusion of proxy variables (e.g., education index or public health expenditure) helps but does not fully resolve this gap.

Third, certain features such as Political Stability or Corruption Perception may suffer from measurement subjectivity or lack of granularity, as they are often based on perception indices rather than hard metrics.

Finally, the dataset covers multiple years, but temporal inconsistencies exist in country-level data availability. Some countries have missing data for certain years, which required imputation or omission, potentially affecting the comparability and generalizability of model outputs.

Recognizing these limitations is essential for interpreting results and guiding future research aimed at refining predictive models and expanding the scope of happiness analysis.

Related Work and How We Compare

A lot of earlier work on happiness prediction mainly used simple models like linear regression, focusing on factors like GDP, social support, and life expectancy (Helliwell et al., 2012). These models are easy to understand but don't always capture the more complicated relationships between factors. More recent studies have moved toward machine learning methods like Random Forests, Gradient Boosting, and SVR because they do a better job with non-linear patterns (Kavakliotis et al., 2018; Tesarova & Benda, 2019). Our project follows a similar path but also tries out Deep Learning to see if it can pick up even deeper trends (Trabucco et al., 2020).

One thing we did differently was include a wider set of features — not just the basics, but also things like Internet Access, Mental Health Index, and Work-Life Balance. This lines up with newer research that highlights the value of using a broader range of indicators for happiness prediction (Munoz et al., 2021). Our results back up what other research has shown: ensemble models like Random Forest and XGBoost are really strong, but deep learning might have even more potential, especially as the data gets bigger and more detailed.

Related Work:

Previous studies have extensively explored various statistical and machine learning approaches to predict happiness from global datasets. The primary source dataset used across these analyses is the World Happiness Report [1], a landmark initiative gathering annual data on worldwide happiness factors.

Early studies predominantly employed traditional regression techniques. Helliwell et al. [2] extensively utilized multiple linear regression to investigate the relationships between happiness levels and factors such as GDP, social support, and healthy life expectancy, establishing foundational benchmarks for predictive modeling of happiness.

Subsequent research has transitioned toward advanced machine learning methods. Kavakliotis et al. [3] compared several predictive models, including Random Forest, Gradient Boosting, and Support Vector Machines (SVM), demonstrating that these methods substantially improved accuracy in predicting happiness from survey data. Similarly, Tesarova and Benda [3] confirmed these findings, reporting superior performance of Random Forest and XGBoost models over traditional regression models in predicting global happiness rankings.

More recently, researchers have introduced deep learning methods to capture complex, nonlinear relationships inherent in well-being datasets. Trabucco et al. [4] implemented deep neural networks to predict happiness, particularly exploiting the intricate social interaction patterns found in social network data. Their results indicated enhanced predictive performance due to deep learning's ability to capture higher-order interactions.

Finally, recent efforts by Munoz et al. [5] expanded predictive models beyond traditional economic and social variables. By incorporating broader features such as internet accessibility, environmental indicators, and mental health factors, their work demonstrated significant predictive improvements and offered a more comprehensive understanding of happiness determinants.

Building upon these prior insights, this study aims to integrate the strengths of both advanced machine learning and comprehensive feature sets to enhance prediction accuracy and interpretability.

Conclusion

This project applied advanced machine learning techniques—Linear Regression, Support Vector Regression (SVR), Random Forest, XGBoost, and Deep Learning Scalar Regression—to predict happiness scores from socio-cultural and political indicators. After careful data cleaning, feature engineering, scaling, and correlation analysis, models were trained and evaluated using standardized methods like MAE, MSE, and R^2. The machine learning models, methods like Random Forest and XGBoost, outperformed simpler models, capturing complex non-linear relationships between factors and happiness. Deep learning models showed promising results, especially for capturing deeper patterns that traditional methods may miss. Overall, this project highlights how machine learning can reveal nuanced insights into well-being, supporting more informed and effective policymaking.

References

[1] World Happiness Report, "World Happiness Data Set," World Happiness Report, 2024. [Online]. Available: https://worldhappiness.report/data-sharing/. [Accessed: Apr. 29, 2025].

[2] J. F. Helliwell, R. Layard, and J. Sachs, *World Happiness Report*, New York, NY, USA: The Earth Institute, Columbia Univ., 2012.

[3] I. Kavakliotis, E. Michailidou, D. Tsolis, and A. Drosou, "Machine learning models for happiness prediction," in *Proceedings of the 12th International Conference on PErvasive Technologies Related to Assistive Environments*, 2018, pp. 408–415.

[4] M. Tesarova and P. Benda, "Applying machine learning techniques for predicting world happiness rankings," in *International Conference on Artificial Intelligence and Soft Computing*, 2019, pp. 637–648.

[5] L. Trabucco, A. D'Andreagiovanni, and G. Rossi, "Deep learning approaches to happiness prediction in social networks," *Neural Networks*, vol. 126, pp. 134–145, 2020.

[6] T. Munoz, M. Perez-Ortiz, and P. A. Gutierrez, "Beyond GDP: Using new indicators for happiness prediction," *Expert Systems with Applications*, vol. 177, article 114907, 2021.