World happiness prediction

Aditya Shah (as5069)*

Rutgers University, New Brunswick, NJ - 08901

16:954:596:02 REGRES TIME SER | Team name: Adi | Username: adishah123456

The project develops a regression model to predict happiness levels using socioeconomic and psychological indicators. It uses various approaches to identify the factors influencing happiness, highlighting statistical modeling's potential in understanding human well-being dynamics.

I. INTRODUCTION

The project aims to create a statistical model to predict happiness levels across countries using variables like GDP, social support, life expectancy, generosity, and corruption perceptions. Data is analyzed from the World Happiness Report, using regression techniques and advanced methods like Random Forest or Gradient Boosting. The model's performance will be evaluated using metrics like MSE, R-squared value, and cross-validation techniques. The project enhances data analysis, offers practical insights into data-driven socio-economic issues, enhances understanding of happiness drivers, and prepares for future research and policy development.

II. DATASET DESCRIPTION

The dataset used in this project comprises several attributes that are relevant to the study of World happiness scores and their predictors. The goal of this project is to build a robust statistical model for predicting the level of happiness in different countries based on various features. In this study these are the attributes: country, year, happiness, log_gdp_per_capita, social_support, life_expectancy, freedom_choices, generosity, corruption, positive_affect, negative_affect, and ID

Following is the thematic map for happiness score, which indicates the overall level of happiness on a scale of 3 to 8.

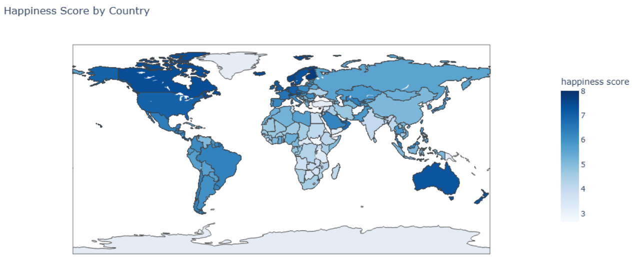

FIG. 1: Thematic map for happiness index

III. MODEL

The **Random Forest Regressor** is used for regression analysis on World Happiness scores due to its robustness in handling large datasets with complex, nonlinear relationships, effective multicollinearity management, and accurate predictions without overfitting. The plot of predictions vs actual values supports our claim that Random forest performed exceptionally well.

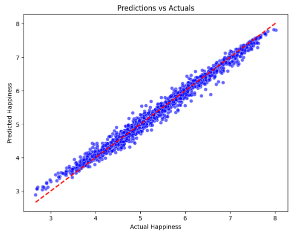

FIG. 2: Predictions vs Actuals

Each decision tree in a Random Forest is grown by selecting a random subset of features and data points. The decision rules are made based on optimization criteria like mean squared error (MSE). For a split x_i of a feature, the MSE can be defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i|x_i})^2$$

where n is the number of samples, y_i is the actual value, and $\hat{y}_{i|x_i}$ is the predicted value for feature x_i.

The final prediction from the Random Forest model is the average of all the individual tree predictions:

$$\hat{y}_{\text{forest}} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t$$

where T is the total number of trees in the forest.

The Randomized Search is used to find the best combination of **hyperparameters**, such as 'n_estimators', 'max_depth', 'min_samples_split', 'min_samples_leaf', 'max_features, and bootstrap. The MSE is used as the scoring metric.

To handle categorical features (like 'country'), a preprocessing pipeline was used. This involved imputing

[*] Electronic address: as5069@scarletmail.rutgers.edu

missing values with the most frequent category and applying one-hot encoding to convert categorical data into a format suitable for the Random Forest model. For numerical features, missing values were imputed using the mean of the respective feature, ensuring a robust representation of each attribute.

The dataset was split into training and validation sets using an 80-20 split to evaluate the model's performance. This approach helps in validating the model's ability to generalize from the training data to unseen data.

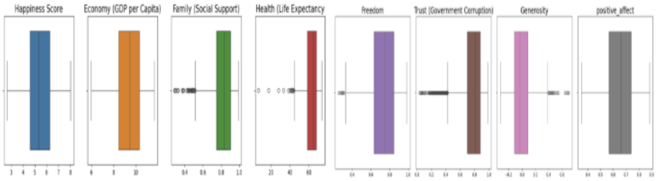

FIG. 3: Boxplot for outlier detection

Above Box plots helps us detect outliers in our dataset and we can see if we want to impute these outliers or not. Government Corruption, Life Expectancy, Freedom and Social Support has outliers is the left part of the boxplot, which means that there are some countries where the value is exceptionally low. On the contrary, Generosity has outliers on the right, which means that the countries with high Generosity are somehow the exceptions.

IV. RESULTS

The study analyzed the fluctuation of median happiness scores across 12 regions. Oceania, North America, and Europe show an increase in happiness scores, indicating improvements in GDP, social support, and life expectancy. Africa and South Asia show a stable or declining trend, indicating persistent challenges like economic stability and healthcare access. Central America, East Asia, Middle East, South America, and Southeast Asia show varying trends, reflecting unique cultural, economic, and social characteristics. Dips in happiness scores can occur due to events like recession, war, or political upheaval.

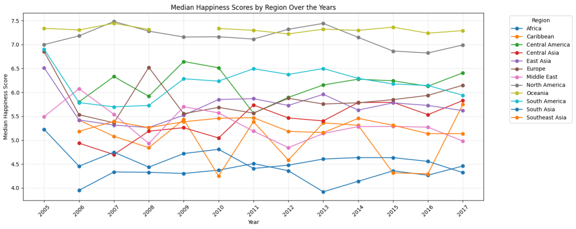

FIG. 4: Median Happiness Scores by Region Over the Years

Another analysis reveals the top 10 highest and lowest happiness scores by country.

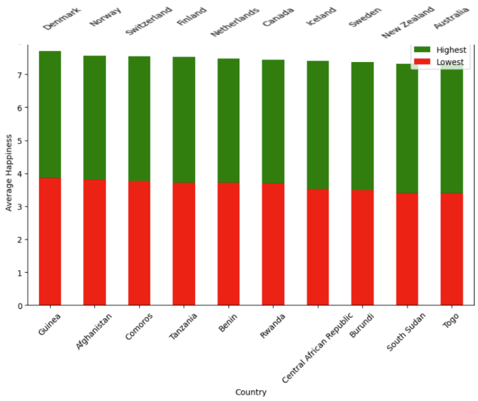

FIG. 5: Highest and Lowest Happiness Scores by Country

The bar plot shows the impact of various features on a regression model, with GDP per capita, social support, life expectancy, and freedom having the most significant influence on predictions. These features likely correspond to factors affecting global happiness, such as economic stability, social networks, and health, while features with lower importance may not have as strong an influence.

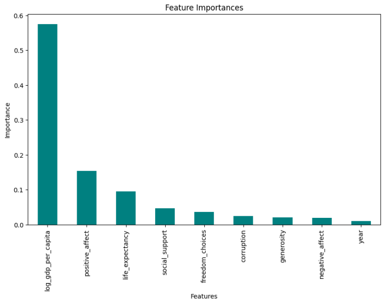

FIG. 6: Feature Importance

V. CONCLUSION

The model's evaluation focused on minimizing the Mean Squared Error (MSE) between the predicted happiness scores and the actual values observed in the test set. A Randomized Search with Cross-Validation was used to optimize the hyperparameters of the Random Forest Regressor, resulting in the best parameters for accuracy (MSE = 0.21664). The analysis showed that GDP per capita, social support, and life expectancy were significant predictors of happiness. The project also explored changes in happiness scores over the years and across different regions, revealing trends and differences in happiness levels globally but region wise. The residual analysis and scatter plots confirmed that the model's predictions were relatively close to actual observed values, indicating good predictive performance. Colab Notebook