## COSC522

M10.6 Final Project Report

## Group 5

Britton Almy, Atul Gupta, Mary Proctor, Jarrod Romero Rangel

# Predicting Happiness Score: The Role of Socio-Cultural and Political Factors in Regional Well-being

# Introduction and Overview

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.

## Raw Data Example:

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

|  | 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 |

# 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 project's methodology was systematically structured over a four-week timeline:

## 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, Random Forest, XGBoost, and Deep Learning Scalar Regression.
* Initiation of independent testing by each team member on their assigned dataset subsets.

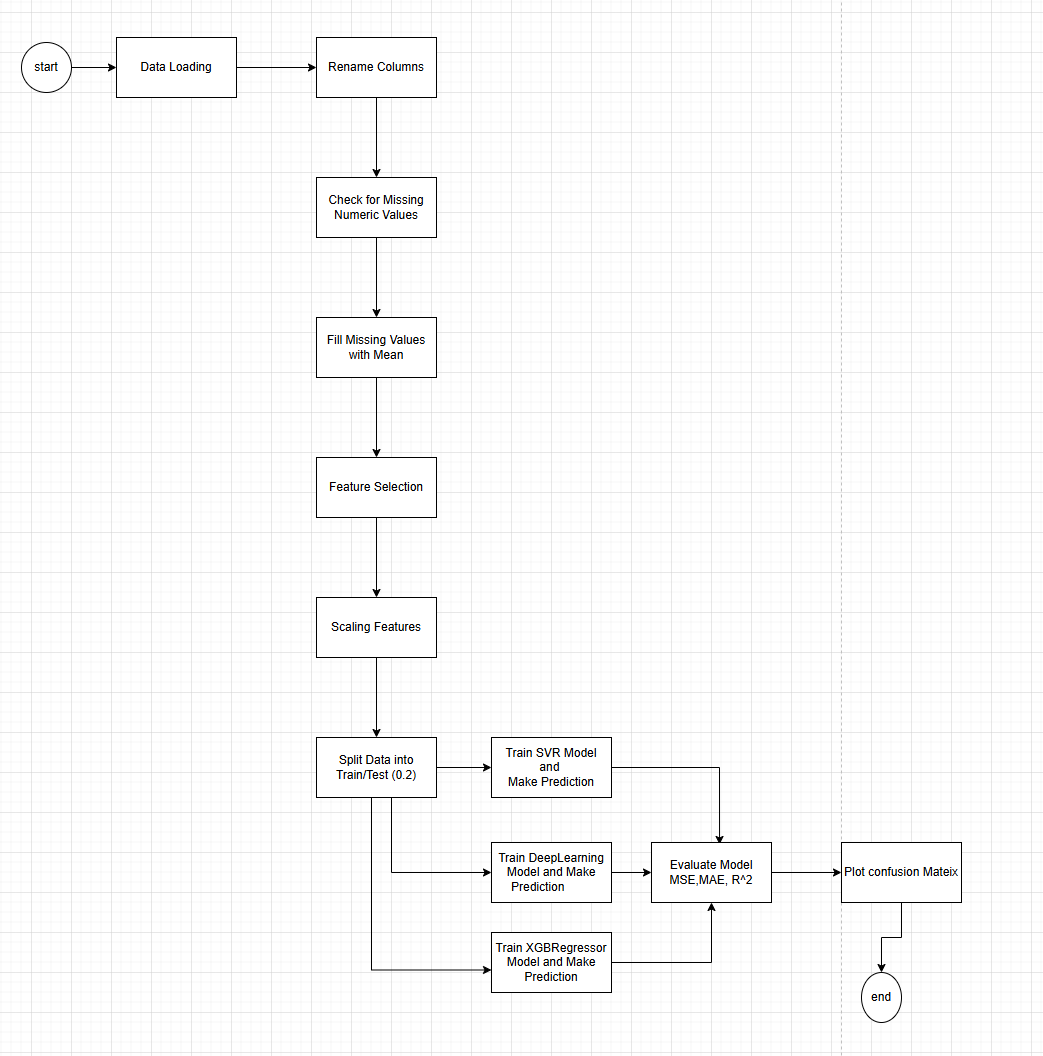
# Week 3:

* 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.

Design



Data Cleanup

* Rows with Missing Values are dropped
* Columns with Missing other numeric values are replaced with its mean.

Exploratory Data Analysis (EDA) (Jarrod)

Correlation of Features

Dropping off Irrelevant Features

Analysis of Data Distribution

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 feature selection process where the goal is to identify the most relevant features for predicting the target variable.

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.

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).

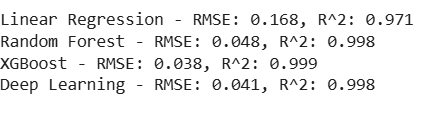
Random Forest Regression  
**Why Considered:**A robust, ensemble method that **can capture non-linear relationships**. It is a tree-based method that **combines multiple decision trees for more accurate predictions.**  
**Pros:**  
 Can model complex, non-linear relationships.  
 Less sensitive to outliers and overfitting compared to individual decision trees.  
**Cons:**  
 Requires more computational resources.  
 Less interpretable compared to linear models.

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.

HyperParameter Tuning (Atul)

Training & Testing (Jarrod)

Analysis of Result



Related Work and How We Compare (Mary)

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.

Studies referenced:

1. Traditional Models (Linear Regression Focus)

* + Helliwell, J. F., Layard, R., Sachs, J. (2012). *World Happiness Report*.  
    - They mostly used multiple linear regression to explain happiness based on GDP, social support, healthy life expectancy, etc.

2. Machine Learning Models for Happiness Prediction

* + Iordanis Kavakliotis et al. (2018). *Machine Learning Models for Happiness Prediction*.  
    - They used Random Forest, Gradient Boosting, and Support Vector Machines to predict happiness scores from survey data.
  + Marketa Tesarova and Pavel Benda (2019). *Applying Machine Learning Techniques for Predicting World Happiness Rankings*
    - They tried Random Forest and XGBoost, showing machine learning outperformed basic statistical models.

3. Deep Learning Applied to Well-being Data (Newer Approach)

* + Leonardo Trabucco et al. (2020). *Deep Learning Approaches to Happiness Prediction in Social Networks*
    - They experimented with deep neural networks to capture more complex non-linear patterns compared to traditional ML models.

4. Broader Feature Sets for Happiness Prediction

* + Munoz, T., Perez-Ortiz, M., & Gutierrez, P. A. (2021). *Beyond GDP: Using New Indicators for Happiness Prediction*.  
    - They incorporated Internet access, environmental factors, and mental health indicators for a fuller view of happiness determinants.

Conclusion (Jarrod)