A Project Report
on
Happiness Prediction: A Statistical Learning Approach

By
Rabin Nepal
(U00901360)
Department of Electrical and Computer Engineering

Submitted to
Baris Kopruluoglu, Ph.D.
Department of Mathematical Sciences

UNIVERSITY OF MEMPHIS
MEMPHIS, TN, USA

December 8, 2023

1 Introduction

Happiness is a term often used but rarely defined with precision. The Oxford Dictionary defines happiness as *"embodies the state of experiencing pleasurable contentment"* [1]. But in practice, it is a more abstract and subjective emotion that encompasses feelings of joy, satisfaction, and overall well-being. Despite its speculative nature, happiness plays a pivotal role in shaping human experiences and is regarded as a fundamental aspect of a fulfilling life.

Motivated by its profound impact on individuals and societies, researching happiness involves delving into diverse factors, ranging from psychological and social factors to cultural and environmental influences. The study of happiness not only aims to understand the influential factors of happiness but also seeks to unravel its complexities, offering insights that can contribute to well-being, policy-making, and broader societal understanding.

In essence, while challenging to quantify, happiness remains a crucial aspect of human existence, driving research, introspection, and initiatives geared toward enhancing individual and collective life satisfaction. In fact, the pursuit of happiness is a universal human endeavor.

1.1 Problem Statement

Understanding the intricacies of happiness remains a significant challenge due to its subjective and multifaceted nature. It encompasses a range of emotions and perceptions, varying greatly among individuals and influenced by a multitude of internal and external factors. Despite efforts to understand the factors contributing to happiness, accurately forecasting an individual's happiness level remains a complex endeavor.

This project attempts to explore various elements influencing happiness and develop a model capable of accurately predicting a happiness score.

1.2 Objectives

The objectives of this project are:

- **Identify Key Factors:** Investigate and uncover influential factors affecting happiness levels.
- **Develop Predictive Models:** Build a robust and reliable model capable of accurately predicting individual happiness scores based on identified factors.
- **Promote Well-being:** Leverage insights derived from the data analysis and predictive modeling to contribute meaningfully towards improving overall well-being and contentment.

2 Dataset

There are many happiness datasets available that can be used for happiness prediction. The one used for this project is the World Happiness Report Dataset. This dataset is specifically chosen because of its global scope, comprehensive nature, and most importantly, it is freely accessible to the public.

This dataset is itself a compilation of survey/interview data from the Gallup World Poll which tracks human development worldwide. The poll focuses on understanding the hopes, dreams, and behaviors of people across the globe. It tracks various aspects of life, including food access,

employment, leadership, well-being, satisfaction, etc. Sustainable Development Solutions Network (SDSN), a non-profit organization under the United Nations (UN) compiles the data from the Gallup World Poll and creates a comprehensive summary of happiness levels in different countries of the world. This report was started in 2012 and is ongoing and includes data on various factors considered to contribute to happiness, such as income, social support, life expectancy, freedom, generosity, etc. The variables used for prediction (predictors) and the response have been explained more below: [2]

- **GDP per Capita:** It represents the country's Gross Domestic Product (GDP) divided by its population, reflecting the average economic output per person. It serves as an indicator of a nation's wealth and standard of living. The GDP per capita is measured in USD and later log scaled for better representation by the WHR dataset.
- **Social Support:** This variable considers factors like family support, community involvement, and access to supportive relationships or simply having someone to count on in times of trouble.
- **Life Expectancy:** Life expectancy represents the average number of years a person is expected to live at birth in a specific country or region. It's a fundamental indicator of the overall health and well-being of a population.
- **Freedom:** This parameter assesses the degree of political and individual freedom within a society. It includes factors such as civil liberties, political rights, and the absence of oppressive conditions.
- **Generosity:** Generosity reflects the willingness of individuals within a society to engage in charitable acts, donate money, or help others without expecting anything in return.
- **Corruption:** Corruption measures the perceived levels of corruption within a country, considering factors like bribery, the integrity of public institutions, and the trustworthiness of governmental bodies.

2.1 Data Pre-processing

This project utilized a large dataset covering the years 2015 to 2023. Although the number of variables varied from 9 to 20 depending on the year, all years shared at least 8 common variables: Country, Happiness Score, GDP per Capita, Social Support, Life Expectancy, Freedom, Generosity, and Corruption.

To prepare the data for analysis, the following steps were taken:

- **Filtering:** The dataset was filtered to include only the 8 common variables shared by all years.
- **NA Removal:** Any missing values (NAs) were removed from the remaining data.
- **Combining Data:** Data from different years were combined using the rbind() function in R.
- **Scaling:** All numerical variables except the response "Happiness Score" were standardized, where the mean and standard deviation for the scaled variables become 0 and 1 respectively.

- **Train-Test Split:** The final dataset was then split into training and testing sets with a ratio of 9:1. This split ensured a sufficient amount of data for model training and evaluation.
  - Training data size: 1228 rows and 7 columns.
  - Test data size: 137 rows and 7 columns.

The original dataset is available at [3]. The preprocessed dataset is also available in Kaggle at [4], [5], and [6]. For this project, the original dataset was studied processed, and made available at [7].

2.2 Exploratory Analysis

The exploratory analysis phase involved a comprehensive examination and understanding of the World Happiness Report dataset. Various statistical techniques, visualizations, and data summarization methods were employed to gain insights into the dataset's characteristics, distributions, correlations, and trends.

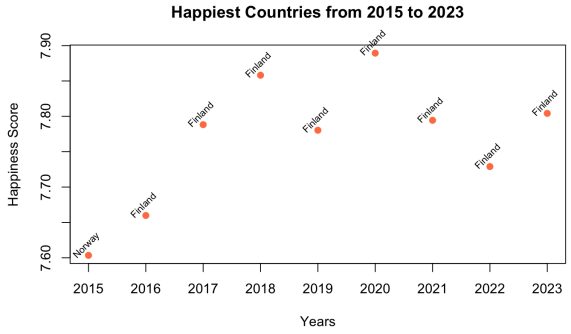

Figure 1: Trend of Highest Happiness Scores: Finland has been the happiest country more consistently

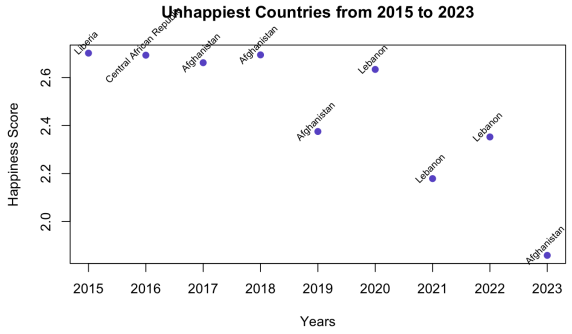

Figure 2: Trend of Lowest Happiness Scores: Afghanistan has been the least happiest country more consistently followed by Lebanon.

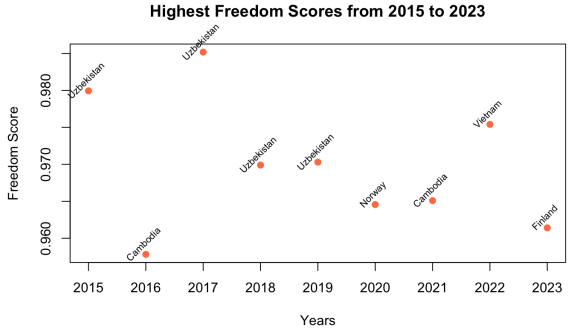

Figure 3: Trend for Highest Freedom Scores: Uzbekistan has consistently been the most free country during the years of study. However, Finland ranked the highest in 2023.

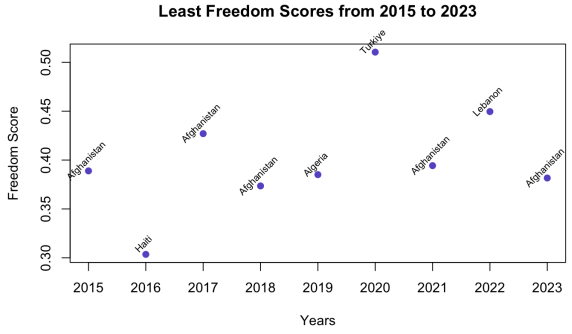

Figure 4: Trend for Lowest Freedom Scores: Afghanistan is the most restricted country in 2023 and it has been so, consistently over the previous years of study too.

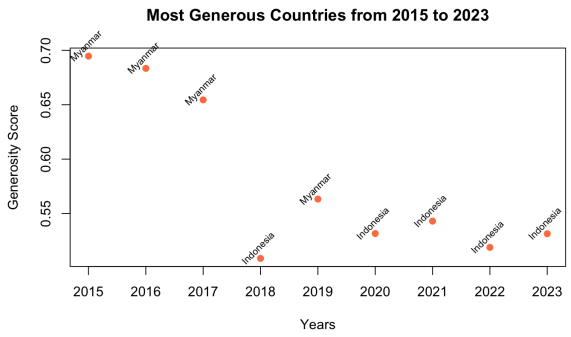

Figure 5: Trend for Highest Generosity Scores: Indonesia and Myanmar have shared the title place of the most generous country over the years.

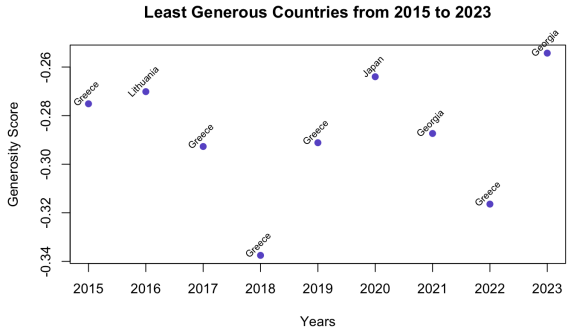

Figure 6: Trend for Lowest Generosity Scores: Greece has consistently ranked the least generous country over the year. In 2023, however, Georgia is ranked the least generous.

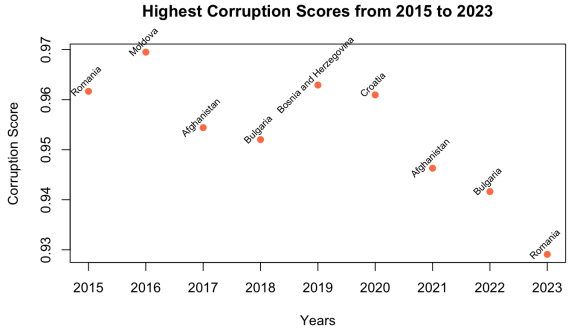

Figure 7: Trend for Highest Corruption Scores: Among all the countries, citizens of Romania have the highest perception that their Government is corrupt in 2023.

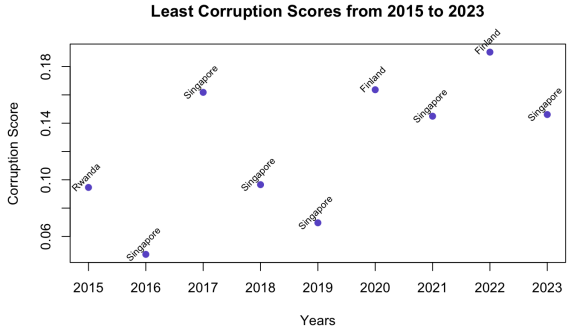

Figure 8: Trend for Lowest Corruption Scores: In 2023 and in most of the previous year's observations too, citizens of Singapore have the least perception of corruption in their government.

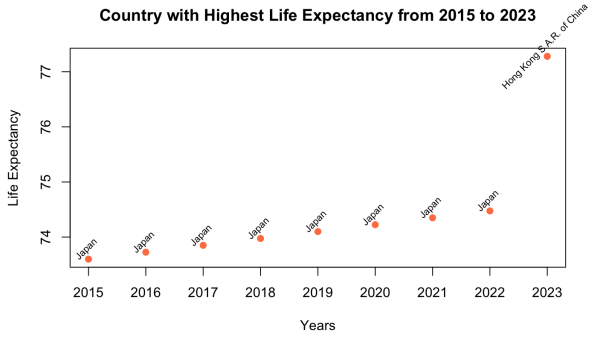

Figure 9: Trend for Highest Life Expectancy: Japan had been leading the highest life expectancy metric (centered around 74 years), but in 2023 Hong Kong ranked first in Life Expectancy with a value of 77.52 years

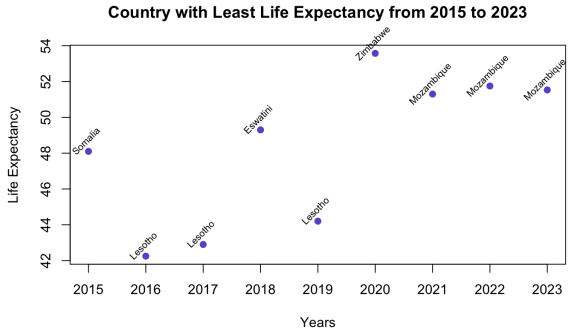

Figure 10: Trend for Lowest Life Expectancy: African countries Zimbabwe and Mozambique appear to have the lowest life expectancy with a value ranging from 52 to 55 years.

3 Approach

The happiness prediction task is a regression problem since the response variable is a quantitative variable that ranges from 0 to 10, where a value close to 0 means the unhappiest and a value

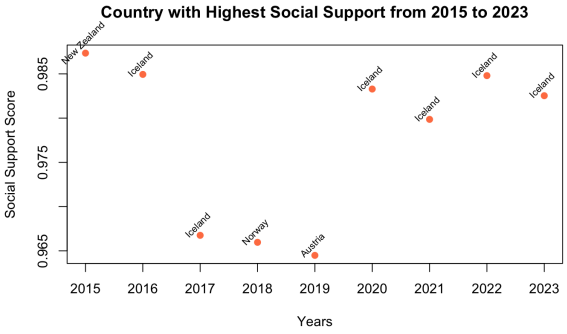

Figure 11: Trend for Highest Social Support Scores: Iceland is consistently the country with the most supportive relationships in its country.

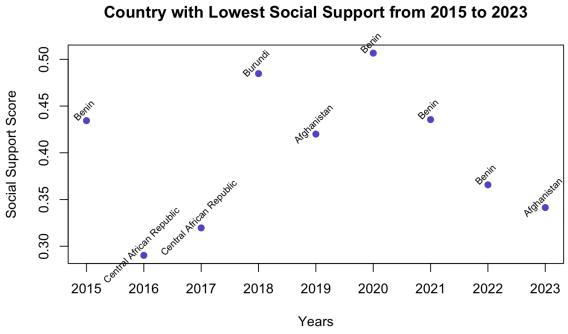

Figure 12: Trend for Lowest Social Support Scores: Benin and Afghanistan have had more social issues than any other country since 2019.

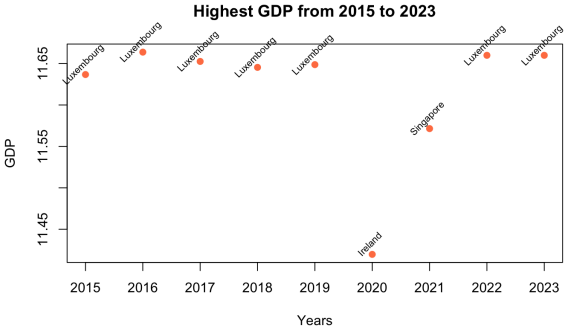

Figure 13: Trend for Highest GDP: Luxembourg is consistently the highest revenue per capita generating country in the world.

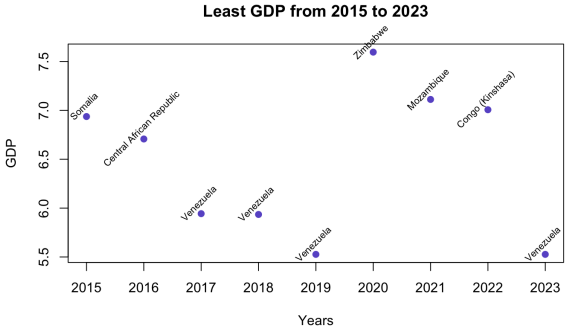

Figure 14: Trend for Lowest GDP: Venezuela is among the most poorest country in the world, ranking lowest in 2017, 2018, 2019 and 2023.

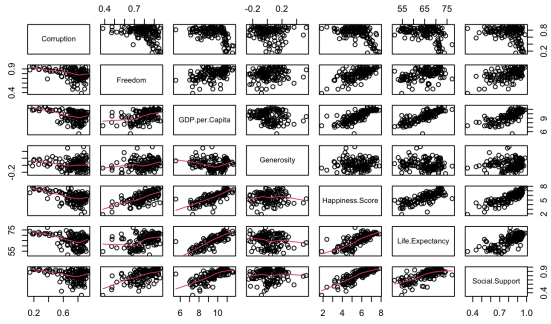

Figure 15: Correlation Plot for all six predictors with the response (happiness score)

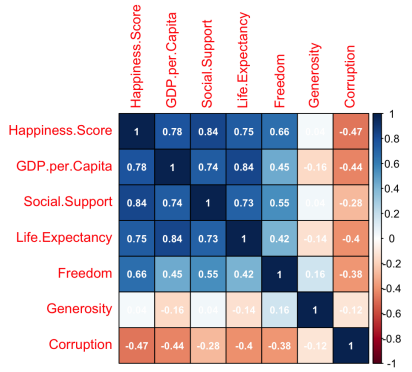

Figure 16: Correlation Coefficients for all variables

close to 10 means the happiest. Consequently, different regression models will be used to make this happiness score prediction. Also, all the predictors (GDP per Capita, Corruption, Freedom, Social Support, Life Expectancy, and Generosity) used for the regression task are quantitative. The regression models used for the prediction task have been explained below:

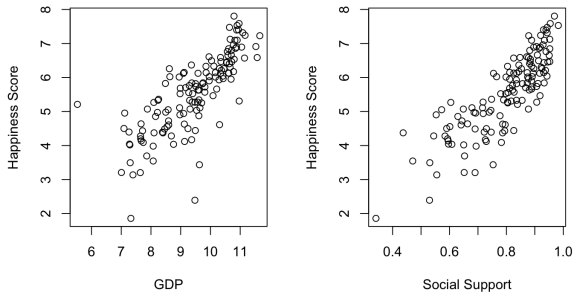

Figure 17: Correlation plot for GDP and Social Support: Both the predictors are significant positive association with the response variable

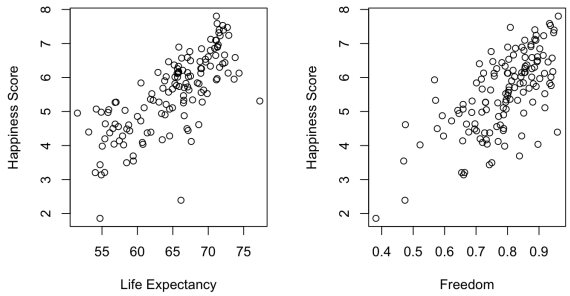

Figure 18: Correlation plot for Life Expectancy and Freedom: Both the predictors are significant positive association with the response variable

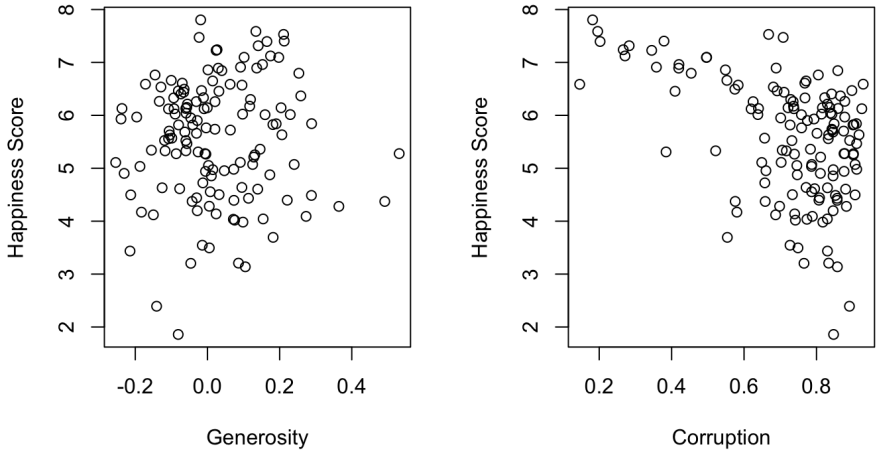

Figure 19: The predictors Generosity and Corruption have been visualized in the plot above. It is apparent that corruption has a significant negative association with happiness, but the predicted does not appear to be correlated to happiness score (correlation coefficient of 0.04).

3.1 Linear Regression Model

A linear regression model is a statistical method that attempts to model the linear relationship between a dependent variable and one or more independent variables by fitting a straight line to the data. The model estimates the coefficients of the line, which can be used to predict the value of the dependent variable for any given value of the independent variable(s). [8]

It is represented mathematically as:

$$Y = \beta_o + \beta_1 * X_1 + \beta_2 * X_2 + ... + \beta_n * X_n + \epsilon \tag{1}$$

where,
Y: response variable
$X_1, X_2, ..X_n$: predictors
$\beta_1, \beta_2, ..., \beta_n$: coefficients of predictors
ϵ: error term

3.2 Lasso Regression Model

Lasso regression, also known as L1-regularized regression, builds on linear regression by adding a penalty term based on the absolute value of the coefficients. This penalty shrinks the less important coefficients towards zero, performing both variable selection and regularization in one step. This leads to a sparser model with fewer features, potentially improving its interpretability and reducing the risk of overfitting. [8]

The penalty term for Lasso is given by:

$$PenaltyTerm = \lambda * \sum_{j=1}^{p} |\beta_j| \tag{2}$$

where,
λ: tuning/regularization parameter
$|\beta_j|$: absolute value of the coefficients

3.3 Ridge Regression Model

Ridge regression, like lasso regression, builds on linear regression by adding a regularization term to the model, but instead of shrinkage towards zero, it shrinks coefficients towards each other. This leads to a smoother model that reduces the variance and improves model generalizability at the cost of potentially losing some information about individual features. [8]

The penalty term for Ridge is given by:

$$PenaltyTerm = \lambda * \sum_{j=1}^{p} \beta_j^2 \tag{3}$$

where,
λ: tuning/regularization parameter
β_j^2: squared magnitude of the coefficients

3.4 KNN Regression Model

K-Nearest Neighbors (KNN) regression predicts the value of a new data point based on the values of its k-nearest neighbors in the training data. It averages the values of those neighbors, making it a simple and efficient regression approach. [8]

4 Results

First, a multiple linear regression model was trained on the training data, and the output of the model was observed. The model showed all the predictors were statistically significant at a 95 percent confidence interval since the p-values for all the coefficients were very small. However, the adjusted R^2 value was observed to be very low at 0.62. The training and test Mean Squared Error (MSE) for the linear regression model were found to be 0.36 and 0.41 respectively.

One of the main objectives of the project is also look into the most influential factors that affect happiness. If there exists some relationship among the predictors themselves, this can

hinder identifying the most influential predictors. A multicollinearity test using Variance Inflation Factors(VIF) was done to see if there exists any correlation among the predictors. As can be seen in Figure 20, the VIF values for GDP per capita and Life Expectancy were observed to be 80.04 and 80.70 respectively, indicating a significant level of multicollinearity among the predictors.

To address the issue of multicollinearity, the predictors- GDP per Capita and Life Expectancy were removed from the linear regression model one at a time, and the output of the model was observed. Analyzing the VIF values after removing the GDP per capita showed that the regression model no longer has highly correlated predictors. A similar observation was made when the predictor Life Expectancy was removed. Both the models, however, performed worse in the prediction task as the adjusted r-squared values were observed to be 0.52 and 0.54 for the models respectively. The training and test MSE were found to be 0.46 and 0.57 for the regression model with GDP per capita removed, and 0.45 and 0.54 for the regression model with Life Expectancy removed.

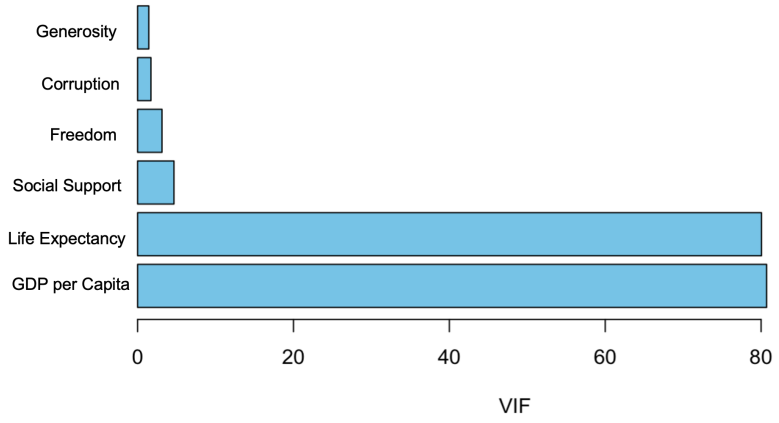

Figure 20: Horzontal Bar plot showing Variance Inflation Factors (VIF) for predictor variables: Higher VIF values of GDP per capita and Life Expectancy indicate a very high correlation among the predictor variables, potentially impacting model accuracy due to multicollinearity.

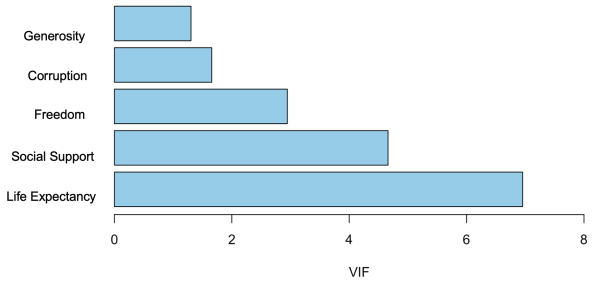

Figure 21: Bar Plot showing VIF for all predictors without GDP per capita in regression model: VIF values range from 1.8 to 7 indicating low to moderate correlation among the remaining predictors. It shows how greatly GDP per capita affects other predictors.

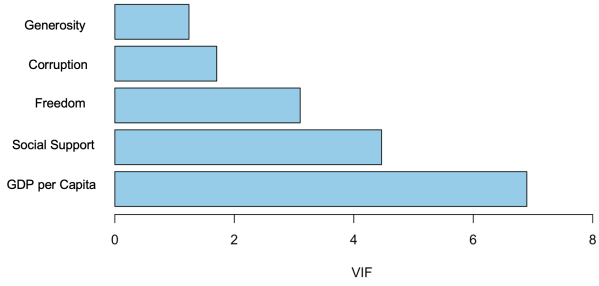

Figure 22: Bar Plot showing VIF for all predictors without Life Expectancy in regression model: VIF values range from 1.8 to 7 indicating low to moderate correlation among the remaining predictors. It shows how greatly Life Expectancy affects other predictors.

Alongside, manually removing predictors from the model, shrinkage methods like Ridge and Lasso regression were also performed to see the change in the output of the model. First, the

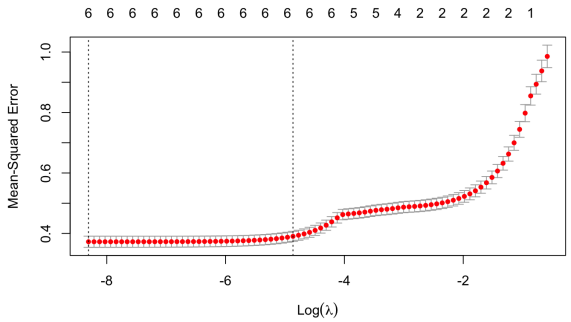

Figure 23: log lambda vs. MSE plot for Lasso Regression

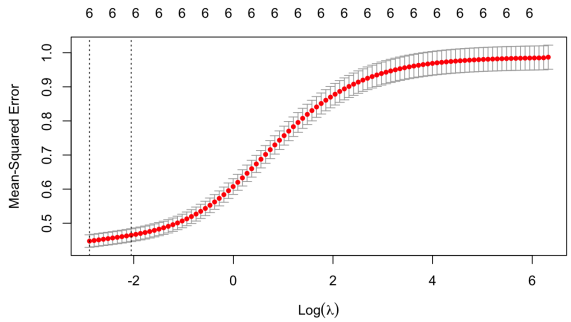

Figure 24: log lambda vs. MSE plot for Ridge Regression

optimum value of the tuning parameter(λ) was determined for both the lasso and ridge regression model using a 20-fold cross-validation method. The value of lambda that minimizes the cross-validation error was found to be 0.00024 and 0.055 respectively for the lasso and ridge regression models. These values of lambda were then used to train the models and the output of the model was observed. Both the models seemed to penalize the coefficients for GDP per Capita and Life Expectancy the most, where the ridge regression imposed a heavier penalty than lasso in this case. In terms of MSE, the lasso regression performed better among the two, obtaining a similar performance to the linear regression model.

Finally, a KNN regression model was trained to predict the happiness score. This model performed the best at N=9 and also performed the best overall with the train and test MSE of 0.16 and 0.26 respectively.

A table summarizing the results of the regression models has been drawn below:

5 Discussion and Conclusion

The analysis conducted in this project indicates various aspects regarding the regression models and predictors in the context of predicting the happiness score. Firstly, it was observed that there exists a significant positive association between the four predictors (Life Expectancy, GDP per capita, Social Support, and Freedom) and the response variable happiness score. The independent variable corruption was the only one that showed a negative association with happiness score. And, Generosity showed no association with happiness score. Secondly, it was observed that there exists significant multicollinearity among the predictors, potentially influencing the stability and accuracy of the linear regression model. Among the models compared, the K-Nearest Neighbors (KNN) model demonstrated superior performance, exhibiting the lowest test Mean Squared Error (MSE) of 0.26, suggesting its higher predictive accuracy compared to other models. Both the Linear Regression and Lasso Regression models displayed similar performance, indicating comparable predictive capabilities for the happiness score. However, the Ridge Regression model yielded a notably worse performance with a higher test MSE than the other models. Moreover, in considering predictor importance, among the six predictors assessed, GDP and Life Expectancy emerged as the most influential factors affecting the happiness score, indicating their higher impact compared to other variables in explaining variations in happiness levels.

To conclude, even though happiness is abstract and subjective, there are factors that can help accurately estimate a score for happiness with some degree of error.

Table 1: A summary of results for the different models

Models	Coefficients		Adj R^2	Train MSE	Test MSE
	Predictors	**Values**			
Linear Regression (all predictors)	Intercept	-0.009296	0.62	0.36	0.41
	GDP per Capita	2.800394			
	Social Support	0.427926			
	Life Expectancy	-2.607748			
	Freedom	0.407279			
	Generosity	0.116610			
	Corruption	-0.188412			
Linear Regression (without GDP)	Intercept	-0.00973	0.52	0.46	0.57
	Social Support	0.54077			
	Life Expectancy	0.08201			
	Freedom	0.53982			
	Generosity	0.04561			
	Corruption	-0.20577			
Linear Regression (without Life Expectancy)	Intercept	-0.01048	0.54	0.45	0.54
	GDP per Capita	0.31934			
	Social Support	0.55548			
	Freedom	0.45269			
	Generosity	0.09745			
	Corruption	-0.31743			
Lasso Regression ($\lambda = 0.000246$)	Intercept	-0.009310632	-	0.36	0.42
	GDP per Capita	2.739552376			
	Social Support	0.430593009			
	Life Expectancy	-2.548467059			
	Freedom	0.408979448			
	Generosity	0.115096837			
	Corruption	-0.188347950			
Ridge Regression ($\lambda = 0.05560$)	Intercept	-0.01019466	-	0.44	0.53
	GDP per Capita	0.37695171			
	Social Support	0.51691395			
	Life Expectancy	-0.21050458			
	Freedom	0.43880235			
	Generosity	0.07401735			
	Corruption	-0.18689692			
KNN (N=9)	-	-	-	0.16	0.26

6 Future Work

This project lays the foundation for further research and development in the field of happiness prediction. The different areas that could be explored are:

- **Examining Temporal Trends:** Analyzing how the influence of various factors on happiness has evolved over time can provide valuable insights into the changing dynamics of well-being.
- **Looking into Regional Differences:** Investigating the regional aspects of happiness can offer a deeper understanding of how geographic and cultural contexts shape individual well-being.
- **Personalizing Happiness Prediction:** Developing personalized models that incorporate individual preferences and experiences holds the potential to significantly improve the accuracy and relevance of happiness prediction.

References

[1] "Oxford english dictionary. happiness. in oxford english dictionary online.." https://www.oxfordlearnersdictionaries.com/us/definition/english/happiness?q=happiness, 2022. [Accessed 07-12-2023].

[2] J. F. Helliwell, R. Layard, J. D. Sachs, L. B. Aknin, J.-E. De Neve, and S. Wang, eds., *World Happiness Report 2023 (11th ed.)*. Sustainable Development Solutions Network, 11 ed., 2023.

[3] SDSN, "Home — worldhappiness.report." https://worldhappiness.report/. [Accessed 07-12-2023].

[4] "World Happiness Report — kaggle.com." https://www.kaggle.com/datasets/unsdsn/world-happiness. [Accessed 07-12-2023].

[5] "World Happiness Report 2015-2021 — kaggle.com." https://www.kaggle.com/datasets/mathurinache/wo happiness-report-20152021. [Accessed 07-12-2023].

[6] "World Happiness Report 2023 — kaggle.com." https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2023. [Accessed 07-12-2023].

[7] R. Nepal, "World Happiness Dataset (2005 to 2023) — kaggle.com." https://www.kaggle.com/datasets/rabinnepal/world-happiness-dataset-2005-to-2023/data, 2023. [Accessed 07-12-2023].

[8] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer US, 2021.