Analysing World Happiness Report Using Data Science

DS203 Course Project Report 2022

Sai Aneesh Suryadevara
190100125
Mechanical Engineering
sai.aneesh@iitb.ac.in

N V Sai Gangadhar
190100080
Mechanical Engineering
saigangadhar.n@iitb.ac.in

Atharva Diwan
190100028
Mechanical Engineering
190100028@iitb.ac.in

Sakshi Toshniwal
190070055
Electrical Engineering
190070055@iitb.ac.in

***Abstract*—The purpose of this study is to look into the data gathered for the World Happiness Report 2020. As more governments, companies, and members of civil society utilise happiness indices to guide their policy decisions, the report continues to acquire awareness on a worldwide scale. We aim to understand the fluctuation of numerous parameters among various categories of countries, we have examined the data set using fundamental data science approaches. This will help us comprehend how these elements differ across geographical areas and how they affect the Happiness Score. In order to forecast the Happiness Index and determine which prediction model performs best for this data set, we tried conventional prediction models. The feature selection for our final data mining models can be done using these findings. We have made predictions using these three models: 1. K-Nearest Neighbors; 2. Decision Tree Regression; 3. Random Forest. F1-score and F2-score are used to compare the performance of the models**

Index Terms: Happiness Index

I. Introduction

The World Happiness Report is a seminal analysis of happiness levels around the world. The first report was released in 2012, followed by reports two and three in 2013 and 2015, respectively, and report four in the 2016 Update. At a celebration honouring International Day of Happiness on March 20, the World Happiness 2017, which rates 155 nations based on their happiness levels, was unveiled. As more governments, companies, and members of civil society utilise happiness indices to guide their policy decisions, the report continues to acquire awareness on a global scale.The report also correlates different (quality of life) elements with the articles and rankings of national happiness, which are based on respondents' assessments of their personal lives.

The Sustainable Development Solutions Network (SDSN) is a global project for the United Nations that brings together development professionals, business and civil society leaders, scientists, and engineers to solve problems using solid data. It supports initiatives for finding solutions that show the capacity of scientific and commercial innovation to aid in sustainable development. The United Nations Sustainable Development Solutions Network recently released the 10th edition of the World Happiness Report 2022. 146 nations in total were listed in the report. Finland has been ranked first in the 2022 World Happiness Report for five years running. Afghanistan, which is ranked 146th on the list of unhappiest nations, tops the list. It was initially used in the 2012 World Happiness Report to measure the happiness of nations. The Global Happiness Council, a team of independent academic happiness experts, developed the Happiness Index. Since 2012, this organisation has yearly published the World Happiness Report (WHR). Earlier, there used to be a survey, in which the respondents were asked to rate their happiness on a scale from 0 to 10. The Happiness Index was calculated by averaging the survey results of the respondents.

The Happiness Index is a comprehensive survey instrument that assesses happiness, wellbeing, and aspects of sustainability and resilience. Happiness Index is also known as comprehensive measure of wellbeing.

Happiness index 2018 is defined as following by the SDSN (life ladder): Think of a ladder with steps that are numbered from 0 at the bottom to 10 at the top. The rungs at the top and bottom of the ladder stand for the finest and worst versions of your potential lives, respectively. Which rung of the ladder do you personally feel you are currently on?

The earlier defition of happiness index is: "The weighted rate of respondents reporting "Very happy" or "Quite happy" less the weighted rate of respondents reporting "Not very happy" or "Not at all happy," plus 100, is the definition of the happiness index. Thus, the index has a range of 0 to 200."

The Cantril ladder asks respondents to picture a ladder, with a 10 representing the best conceivable life for them and a 0 representing the worst possible existence. On a scale of 0 to 10, they are then asked to score their own current lifestyles. The 2019–2021 rankings are based on data from nationally representative samples. However, in this work we have not considered the happiness ladder, rather happiness score index has been used.

According to the World Happiness Report given in Wikipedia, the six key factors affecting the happiness index are real GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity and perceptions of corruption.

This report will aid government organisations and societies in drafting public policies that are aimed towards the betterment of happiness of the people in general. This analysis provides the institutions in understanding the factors which will truly contribute to the increase in happiness of people. The predictive models also give insights into the amount of variation of parameters required to achieve desired happiness index in the society.

II. Related Work

Since 2012, Gallup's World Poll has been a primary source of global data behind the life satisfaction rankings released in SDSN's highly publicized World Happiness Report. So far, SDSN has published eight reports. Research has been done extensively in this field. A few of the works in this area are:

- Exploring the Biological Basis for Happiness
- World Happiness Report 2022
- World Happiness Report 2021
- World Happiness Report 2018 - The "happiest" nation in the world, according to the 2018 Happiness Index, is Finland. The UN released the report on March 14th, 2018. The World Happiness Report is a seminal analysis of happiness levels around the world. The Pontifical Academy of Sciences in the Vatican hosted the launch event for the World Happiness Report 2018, which ranks 156 countries by their levels of happiness and 117 countries by the happiness of their immigrants.
- Defining a New Economic Paradigm: The Report of the High-Level Meeting on Wellbeing and Happiness
- Analysis of Happiness Index
- A Comparative Analysis Of The Factors Affecting Happiness Index - This paper discusses uses multi regression to determine to the extent to which the factors affect the happiness index. It also analyses the difference in the happiness of males and females, and determines the relation between income and happiness index

III. Data and Methodology

The data is corresponding to 2022 statistics obtained from "The World Happiness Report". The factors considered in the analysis are real GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity and perceptions of corruption. First, exploratory data analysis is done to understand the data. Later, hypothesis testing provides us the important features that actually affect the happiness index. This insights are provided based on the p-values and correlation matrix. Later, predictive models such as KNN, Decision Tree regression and random forest are used for data mining. A comparison of these models are provided using f1 score, precision, recall and support. Finally our learnings and conclusions are provided.

IV. Experiments and Results

Exploratory Data Analysis:

We have initially analysed the dataset by finding basic information related to the parameters and happiness score. This includes analysing the number of unique values and entropy of values in each columns.

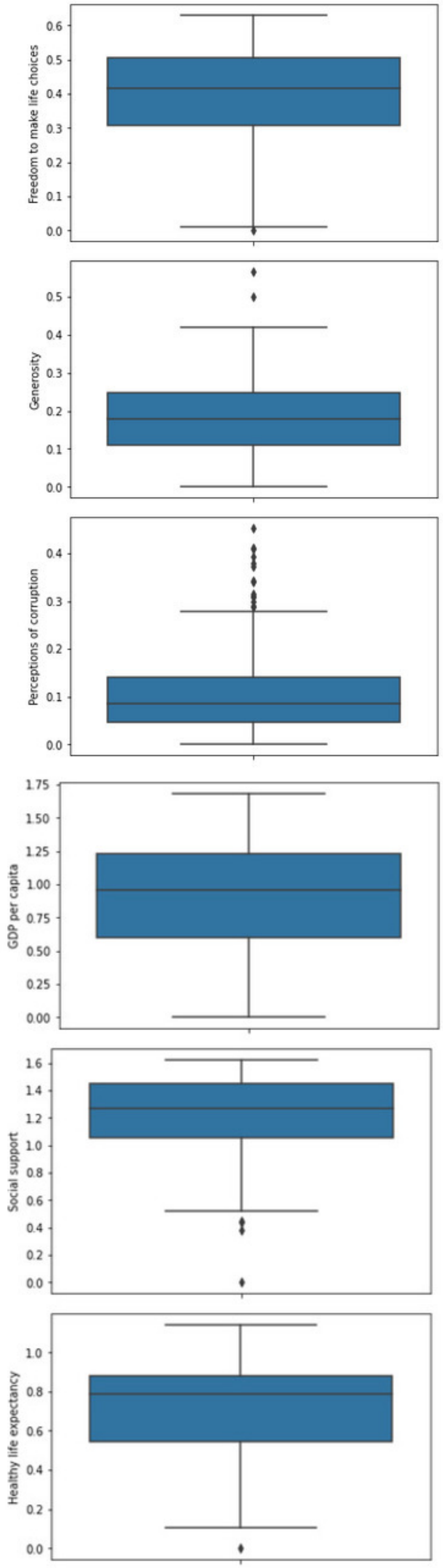


Fig. 1. Box-and-whiskers plot for visualization of quartiles of continuous columns

The variables are categorised based on their types and their statistics are computed. The QQ plots are plotted to understand how much dispersed the dataset is compared to normal distribution. We can see from the Box-and-whiskers plots that there are many anomalies in social support and perceptions of corruption, making them appear to be highly influenced by the opinions of different people.

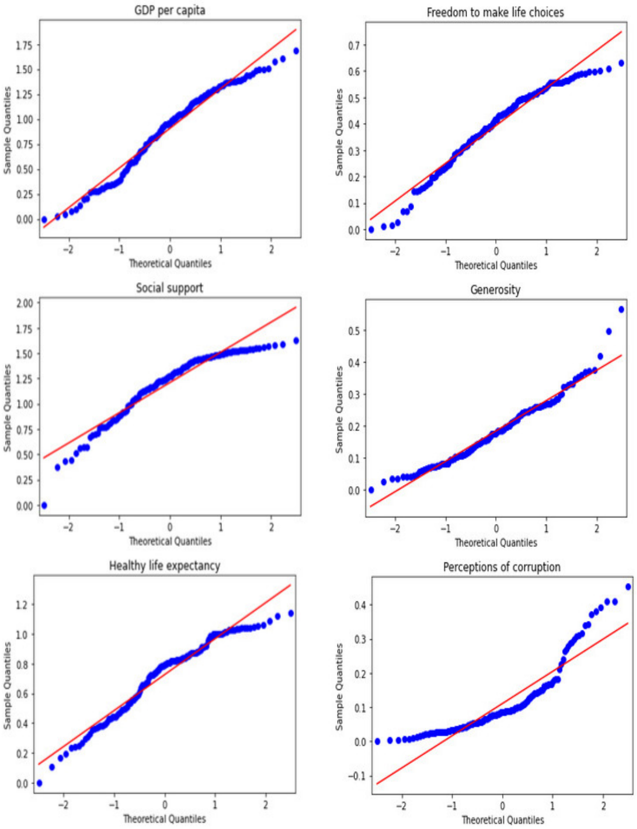


Fig. 2. QQ plots for understanding how much dispersed the dataset is compared to normal distribution

Descriptive Data Analysis: The underlying information in the data collection has been extracted with the help of data analysis. To comprehend the data's variability, box and whisker graphs have been created. The plot demonstrates how social support, perceptions of corruption, and negative affect have a lot of outliers and appear to be quite susceptible to individual viewpoints. To explain the distribution of the data graphically, histograms and Q-Q plots have been employed. The normal distribution is notably different for perceptions of corruption and social support, although it is somewhat followed by the other variables.

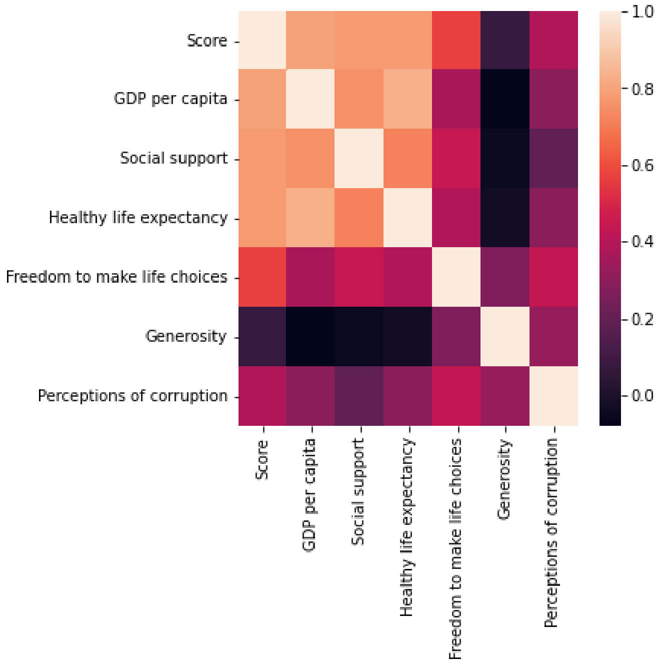
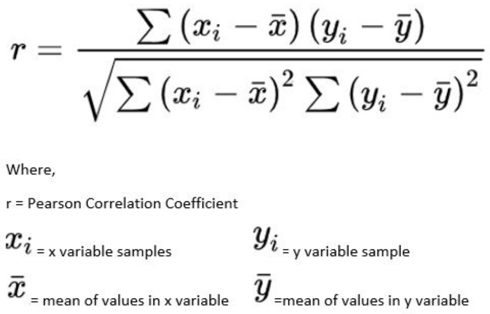


Fig. 3. Correlation matrix for understanding relation between different columns

Hypothesis testing: Here, we statistically examine the relationships between variables and the significance of each variable in calculating the happiness index. To ascertain the degree of correlation between variables, we performed a number of statistical tests. Our target variable-happiness index of country was continuous. We computed **Pearson Correlation** coefficient for continuous variables to determine the linear relationship between variables. Positive Pearson correlation coefficient signifies positive correlation, zero signifies no correlation and negative values signifies negative correlation between variables. We also calculated the p-value to test the dependence of the happiness index on the variable. If the calculated p-value is greater than the significance level we will accept our null hypothesis otherwise reject it.



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

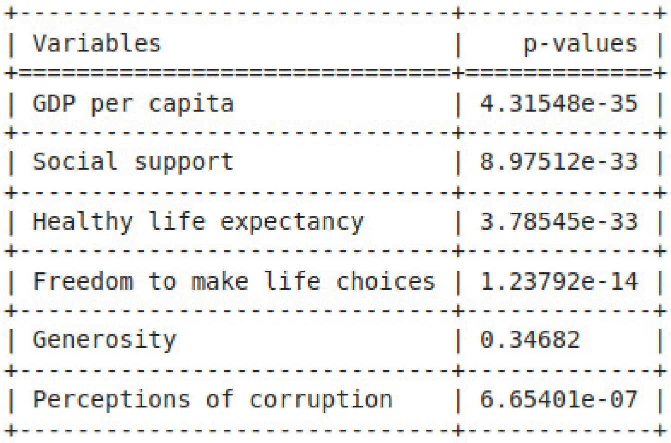
Where,

r = Pearson Correlation Coefficient

x_i = x variable samples y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\bar{y}$ =mean of values in y variable

Fig. 4. Formula for Pearson Correlation Coefficient [8]



```
+------------------------------+-------------+
| Variables                    |    p-values |
+==============================+=============+
| GDP per capita               | 4.31548e-35 |
+------------------------------+-------------+
| Social support               | 8.97512e-33 |
+------------------------------+-------------+
| Healthy life expectancy      | 3.78545e-33 |
+------------------------------+-------------+
| Freedom to make life choices | 1.23792e-14 |
+------------------------------+-------------+
| Generosity                   | 0.34682     |
+------------------------------+-------------+
| Perceptions of corruption    | 6.65401e-07 |
+------------------------------+-------------+
```

Fig. 5. Calculated p-values

We can see that GDP per capita, Social Support, Life expectancy, freedom to make choices an perception to corruption all have very small p-values. So we reject the null hypothesis. However, the p-value of Generosity is very high. This implies we accept the Null Hypothesis. Hence, Happiness index does not depend on the Generosity index of a country. Based on the results we concluded, the happiness index was dependent on the following variables: GDP per capita, Social support, Healthy life expectancy at birth, Freedom to make life choices, and Perceptions of corruption.

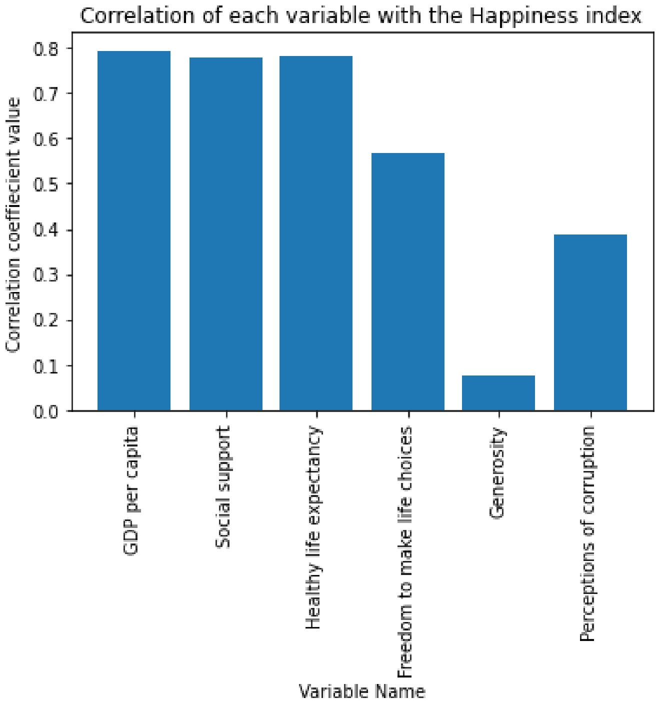


Fig. 6. Correlation of each variable with Happiness Index

The correlation is positive for all these features. Thus we conclude that the happiness index increases with increasing GDP, social support, life expectancy, Freedom to make life choices. But as the negative affect index and perception of corruption in the country decreases the happiness index decreases. This is expected as the negative affect index indicates the level of worry, sadness, and anger people experienced the previous day.

Prediction Models: We first performed data cleaning and preprocessing on the data. We encoded the country names in numerical format using label encoder. The data was split into training and testing data with split ration 75:25. All the Nan values present in the data was replaced with the mean of the values of the variable in the training dataset. Both the training and testing dataset was normalized with a standard scaler fitted on the training dataset. After this, we trained three different data mining models to predict the happiness index of a country. We implemented the following models using scikit learn:

- **K-Nearest Neighbors**
- **Decision Tree Regression**
- **Random Forest Regression**

We utilised **GridSearchCV** from scikit-learn for hyper parameter tuning. Using these parameters, we trained our models. The trained models were used to predict the happiness index of the testing dataset and the accuracy of each of model was recorded.

a)KNN: For this prediction model, the tuned parameter was number of neighbours (the best 'k' value) which was found to be 21. The results from this model are presented below.

Model	Precision	Recall	F1-score	Support	Accuracy
KNN	0.60	0.62	0.60	39	0.62

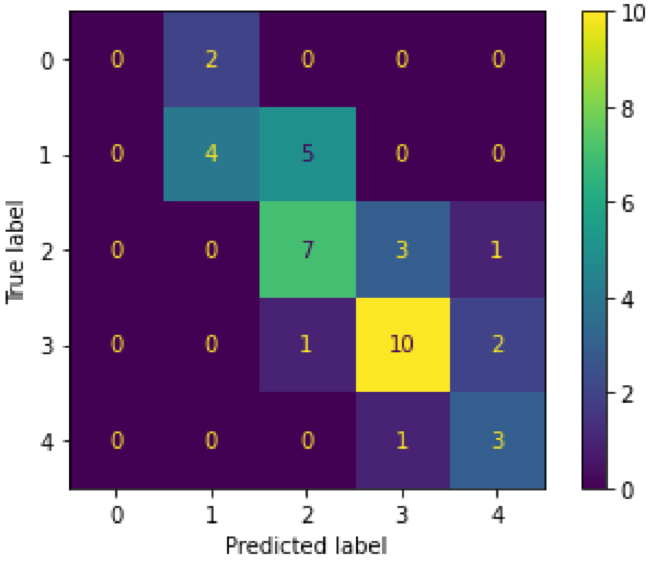


Fig. 7. Comparison of test results

b)Decision Tree: For this prediction model, the there were two hyperparameters to be tuned : 1)Criterion and 2)Max_depth. We have used Gini index and Maximum depth to be 4. The results from this model are presented below.

Model	Precision	Recall	F1-score	Support	Accuracy
Decision Tree	0.59	0.54	0.55	39	0.54

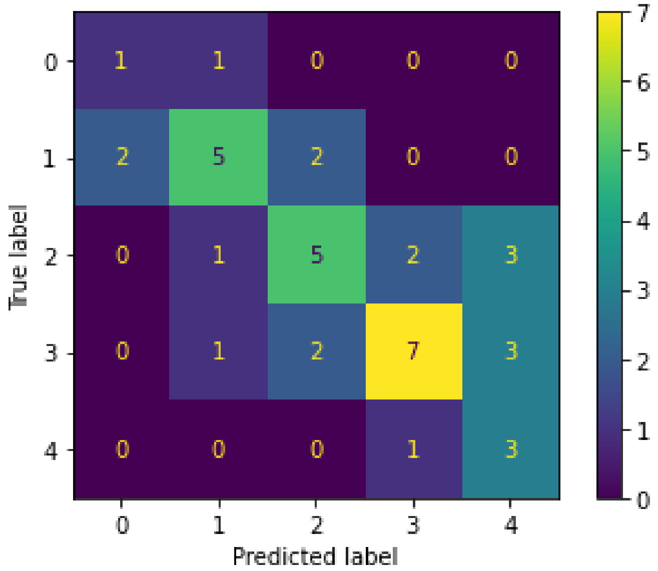


Fig. 8. Comparison of test results

Based on the results, we can see that Decision Tree performs poorly compared to KNN.

c)Random Forests: For this prediction model, the tuned three hyperparameters: 1) Number of estimators 2)Max Depth 3)Model criterion. From GridSearchCV, we used 100 estimators, maximum depth 4 and Entropy as the criterion. The results from this model are presented below.

Model	Precision	Recall	F1-score	Support	Accuracy
Random Forests	0.59	0.59	0.58	39	0.59

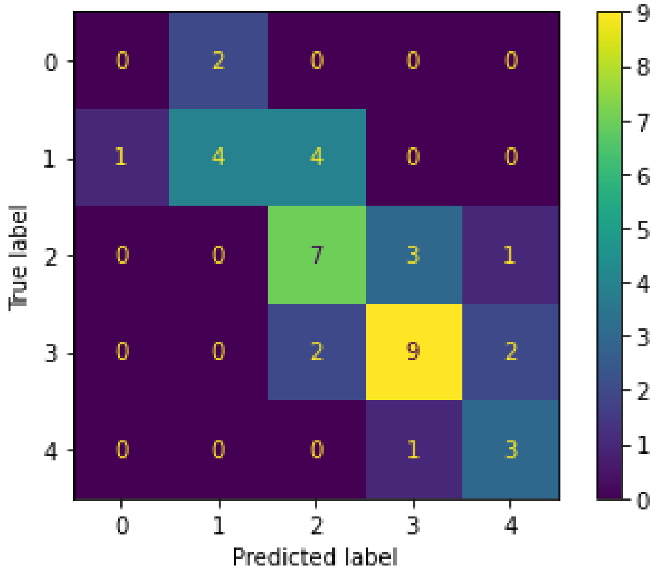


Fig. 9. Comparison of test results

Multiple decision trees are combined by the Random Forest Classifier, improving the model's accuracy and robustness. Consequently, it is more accurate than a decision tree classifier. However, KNN achieved the highest accuracy amongst all the models.

V. Learning and Conclusions

We think that by helping governments prioritise their goals for better nations, our report will benefit them all. The data analysis led to the following conclusions:

- Social support, perceptions of corruption has numerous outliers and appears to be highly influenced by the opinions of different people. It might therefore be more challenging to measure.
- Corruption perceptions and happiness score have very little in common.
- A low Freedom to make life choices seems to make countries miserable, but a high Freedom to make life choices may not always make a country happy. A high GDP per capita is associated with greater happiness.
- A high Social support rating appears to be crucial for achieving a high Score. The coefficient of variance is extremely low in both cases (high and low Social Support), showing that the Scores are consistently high and low for the two cases.
- The Score can be effectively determined by a healthy life expectancy. For both circumstances, the coefficient of variation is small. As a result, one of the easiest methods for governments to raise their Score is to concentrate on health facilities.

Some of the shortcomings of this report are:

- No matter the population of a country, a certain number of individuals are polled there to create the data set. Statistical inconsistencies could result from this.

Contribution of Team Members

- Atharva Diwan and Sai Aneesh did Hypothesis testing and implemented Data Mining models.
- Sakshi Toshniwal did Exploratory Data Analysis and Descriptive Data Analysis.
- Sai Aneesh and Gangadhar read research papers which analysed Happiness Index and made this report.

Acknowledgment

We would like to thank our professors for teaching us this course, without which we would not have been able to complete this project. We would also like to thank all the TAs involved with this course, who helped us a lot in doing the various assignments, and completing this course.

References

[1] World Happiness Report - https://en.wikipedia.org/wiki/World_Happiness_Report.

[2] Meike Bartels, Ragnhild Bang Nes "Exploring the Biological Basis for Happiness" 2022.

[3] Helliwell, J. F., Layard, R., Sachs, J. D., De Neve, J.-E., Aknin, L. B., Wang, S. (Eds.). (2022). World Happiness Report 2022. New York: Sustainable Development Solutions Network.

[4] Helliwell, John F., Richard Layard, Jeffrey Sachs, and Jan-Emmanuel De Neve, eds. 2021. World Happiness Report 2021. New York: Sustainable Development Solutions Network.

[5] “Bhutan’s Gross National Happiness Index,” https://ophi.org.uk/policy/gross-national-happiness-index/.

[6] B. Prashanthi, Dr. R. Ponnusamy, “Future Prediction of World Countries Emotions Status to Understand Economic Status using Happiness Index and SVM Kernel,” International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 11 — Nov 2019.

[7] Yichen Ma, Andrew Liu, Xukai Hu, Yuchen Shao, “Happiness Score Identification: A Regression Approach,” ISEESE 2020.

[8] “Beginner’s Guide to Pearson’s Correlation Coefficient” https://www.analyticsvidhya.com/blog/2021/01/beginners-guide-to-pearsons-correlation-coefficient/

[9] Charles Alba, “A Data Analysis of the World Happiness Index and its Relation to the North-South Divide,” Undergraduate Economic Review, Volume 16 Issue 1, 2019.

[10] Meghna Chaudhary, Siddharth Dixit, Niteesh Sahni, “Network Learning Approaches to study World Happiness,” July 21, 2020.