STUDY OF WORLD HAPPINESS REPORT DATASET USING MACHINE LEARNING METHODOLOGIES

Mahalakshmi A
mahalakshmi.arasu2023@vitstudent.ac.in
M.Sc. Data science, Department of Mathematics, Vellore Institute of Technology (VIT), Vellore, Tamil Nadu, India-632014.

Hariharan R B
hariharan.2023@vitstudent.ac.in
M.Sc. Data science, Department of Mathematics, Vellore Institute of Technology (VIT), Vellore, Tamil Nadu, India-632014.

Kamalesh K
kamalesh.k2023@vitstudent.ac.in
M.Sc. Data science, Department of Mathematics, Vellore Institute of Technology (VIT), Vellore, Tamil Nadu, India-632014.

***Abstract*—The World Happiness Report, an illustrious annual publication, is highly regarded for its role in providing insights into the happiness levels of nations across the globe. While its primary purpose is to assess social well-being, it extends its relevance into the realm of commerce by offering valuable information that can help address significant business challenges. In today's corporate landscape, organizations are increasingly recognizing the profound impact that employee well-being has on overall performance, job satisfaction, and ultimately, the success of their enterprises. Machine learning, including techniques like Simple Linear Regression, Random Forest Regression, and Decision Tree Regression, has emerged as a powerful tool in understanding the intricate link between contentment and workplace efficiency. These advanced analytical methods can help organizations predict and analyze how employee happiness affects their engagement, motivation, and productivity. By employing Machine Learning models, organizations can gain deeper insights into the factors that contribute to employee well-being and, in turn, make more informed decisions to enhance workplace satisfaction. The realization that happy and content employees tend to be more engaged, motivated, and productive has spurred organizations to turn to sources like the World Happiness Report for guidance and leverage data-driven techniques such as Machine Learning to optimize their strategies for fostering a positive work environment and, consequently, improving their overall performance and success.**

Keywords— World Happiness Report, Annual publication, Happiness levels, Machine learning, Simple Linear Regression, Random Forest Regression, Decision Tree Regression.

I. INTRODUCTION

Every year, the World Happiness Report is released, ranking the nations according to their levels of happiness and offering insightful information about the elements that affect people's quality of life. This report dataset comprises a rich collection of data points, including social, economic, and environmental variables, which offer a holistic view of happiness across different nations. Through the application of machine learning methodologies, we can harness the power of data to uncover hidden patterns, relationships, and predictive insights within this dataset. This opens up a wealth of possibilities for research and analysis, enabling us to address questions such as:

1. What are the key determinants of happiness across countries?
2. Can we predict a nation's happiness ranking based on its socio-economic indicators?
3. How do societal and environmental factors impact happiness levels over time?
4. Are there commonalities or differences in the drivers of happiness among nations?

This introduction sets the stage for a comprehensive exploration of the World Happiness Report dataset using machine learning techniques. We seek to learn more about the intricate and varied aspects of happiness in our global society by utilizing data and cutting-edge approaches. Ultimately, our findings and insights can contribute to informed policy decisions and help shape a happier, more prosperous world for all.

A. Objective

In this report, we embark on a data science project that harnesses the power of the World Happiness Report to explore the intricate link between happiness and workplace productivity. Our objective is to uncover the factors that contribute to happiness in different countries and draw connections to the realm of the workplace. By undertaking a comprehensive analysis of the economic, social, and environmental indicators within the report, we seek to offer evidence-based insights that can shape organizational strategies aimed at promoting employee happiness and, consequently, enhancing productivity.

B. Problem definition

The World Happiness Report highlights a critical issue that has to be addressed: the ongoing happiness gap in developing countries. While some countries experience high levels of happiness and well-being, others continue to face significant challenges in achieving similar outcomes. This discrepancy raises important questions about the underlying causes of the happiness gap and its implications for social and economic development. The problem at hand is twofold. Firstly, there is a need to understand the factors contributing to the lower happiness levels observed in developing nations. This requires a comprehensive analysis of economic, social, and environmental indicators, as well as an exploration of the unique circumstances and challenges faced by these countries. Identifying the root causes of the happiness gap is crucial in formulating effective strategies to address this issue. Secondly, it is imperative to recognize the consequences of the happiness gap on the overall well-being and progress of developing nations. Happiness has been found to be closely linked to various indicators of human development, including

health, education, productivity, and social cohesion. Therefore, the persistence of the happiness gap can hinder progress in these areas, perpetuating a cycle of underdevelopment and inequality.

C. Proposed solution

Our project aims to provide a comprehensive solution that empowers users to predict the happiness score of their country and gain valuable insights into the factors influencing happiness. To achieve this, we have developed a user-friendly website that allows individuals to enter their country's data along with their personal information. Leveraging advanced data science techniques, we can generate personalized predictions of happiness scores and offer users a deeper understanding of the contributing factors.

II. LITERATURE SURVEY

[1] **Title**: "World Happiness Report (2023)"
Authors: Andrew N. Rowan, Wellbeing International
Summary: In general, national well-being has been remarkably resilient despite the increase in global morbidity and mortality caused by the pandemic and the Russian invasion of Ukraine. Helping and altruistic behaviors increased during the pandemic and may have acted to counter the adverse effects of the pandemic and associated lockdowns.

[2] **Title**: "A Data Analysis of The World Happiness Index and Its Relation to The North-South Divide"
Authors: Charles Alba
Summary: A primary objective is to demonstrate how qualitative data offering bottom-up perspectives on wellbeing offer a necessary complement to quantitative self-report measures, allowing for more nuanced cultural understandings of lived experience and wellbeing that recognize diversity both globally and locally. The research contextualized responses to standardized life evaluations (including the Cantril ladder question used by the World Happiness Report) through observations and interviews along with culturally sensitive measures of emotional experience.

[3] **Title**: "A Measure of Well-Being Efficiency Based On The World Happiness Report"
Authors: Francesco Sarracino, Kelsey O'Connor
Summary: We estimate a measure of well-being efficiency that assesses countries' ability to transform inputs into subjective well-being. We use the six inputs (real GDP per capita, healthy life expectancy, social support, freedom of choice, absence of corruption, and generosity) identified in the World Happiness Reports and apply Data Envelopment Analysis to a sample of 126 countries. Efficiency scores reveal that high-ranking subjective well-being countries, such as the Nordics, are not strictly the most efficient ones. Also, the scores are uncorrelated with economic efficiency.

[4] **Title**: "Analysing Happiness Index As A Measure Along With Its Parameters And Strategies For Improving India's Rank In World Happiness Report"
Authors: Francesco Sarracino, Kelsey O'Connor
Summary: We estimate a measure of well-being efficiency that assesses countries' ability to transform inputs into subjective well-being. We use the six inputs (real GDP per capita, healthy life expectancy, social support, freedom of choice, absence of corruption, and generosity) identified in the World Happiness Reports and apply Data Envelopment Analysis to a sample of 126 countries. Efficiency scores reveal that high-ranking subjective well-being countries, such as the Nordics, are not strictly the most efficient ones. Also, the scores are uncorrelated with economic efficiency.

[5] **Title**: "World Happiness Report (2012)"
Authors: John Helliwell, Richard Layard and Jeffrey Sachs
Summary: The report, published by the Earth Institute and co-edited by the institute's director, Jeffrey Sachs, reflects a new worldwide demand for more attention to happiness and absence of misery as criteria for government policy. It reviews the state of happiness in the world today and shows how the new science of happiness explains personal and national variations in happiness.

[6] **Title**: "World Happiness Report (2017)"
Authors: John Helliwell, Richard Layard and Jeffrey Sachs
Summary: This report gives special attention to the social foundations of happiness for individuals and nations. The chapter starts with global and regional charts showing the distribution of answers, from roughly 3000 respondents in each of more than 150 countries, to a question asking them to evaluate their current lives on a ladder where 0 represents the worst possible life and 10 the best possible.

[7] **Title**: "Clustering Countries According To The World Happiness Report (2019)"
Authors: M. Mujiya Ulkhaq, Arga Adyatama
Summary: This study explores nine clustering algorithms were presented, and three internal validation indices were utilized to compare the algorithms. k-medoids algorithm was selected to illustrate how three distinguished clusters generated from the algorithm are different from each other. This study is expected to give an insight into how to implement clustering algorithms into the real-world data set and how to interpret the results.

[8] **Title**: " Analysis of World Happiness Report Dataset Using Machine Learning Approaches "
Authors: Moaiad Ahmad Khder, Mohammad Adnan Sayfi and Samah Wael Fujo
Summary: The purpose of this paper is to conduct a study on World happiness report dataset, to classify the most critical variables regarding the life happiness score. The strong evidence of the identified main features classified from the outcomes of applying the supervised machine learning approaches using the Neural Network training model

and the OneR models in classifications and feature selection.

[9] **Title**: "What the World Happiness Report Doesn't See: The Sociocultural Contours Of Wellbeing In Northern Tanzania"
Authors: Michael Kaufman, Andrew M. Guest
Summary: A primary objective is to demonstrate how qualitative data offering bottom-up perspectives on wellbeing offer a necessary complement to quantitative self-report measures, allowing for more nuanced cultural understandings of lived experience and wellbeing that recognize diversity both globally and locally. The research contextualized responses to standardized life evaluations (including the Cantril ladder question used by the World Happiness Report) through observations and interviews along with culturally sensitive measures of emotional experience.

[10] **Title**: "Analysing Happiness Index as A Measure Along With Its Parameters And Strategies For Improving India's Rank In World Happiness Report"
Authors: Sarah Ahtesham
Summary: Measuring happiness in quantifiable terms is a global phenomenon lately. United Nation's World Happiness Report (WHR) is one such means to analyse the level of subjective wellbeing that countries across the world are living with. The Happiness Index is framed to set various parameters on grounds of which a country could be ranked in a list of 156 countries. India's rank has come down the list this year (2019) to be ranked at the 140th position.

III. MATERIALS AND METHODS

A. *Data Collection*

We used Secondary data from sources like Kaggle to examine people's happiness data. It is a common and valuable approach for research and analysis.

B. *Methodology*

We have collected data, including metrics like ladder score, upper whisker, lower whisker, logged GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, perceptions of corruption, dystopia, residual from sources like Kaggle to examine factors that contribute to happiness of people in each country.

We must undergo several essential steps in each research study, like collecting data, pre-processing data, selecting a suitable model, implementing it, calculating errors, and creating results. So, following the step-wise procedure is required as shown below,

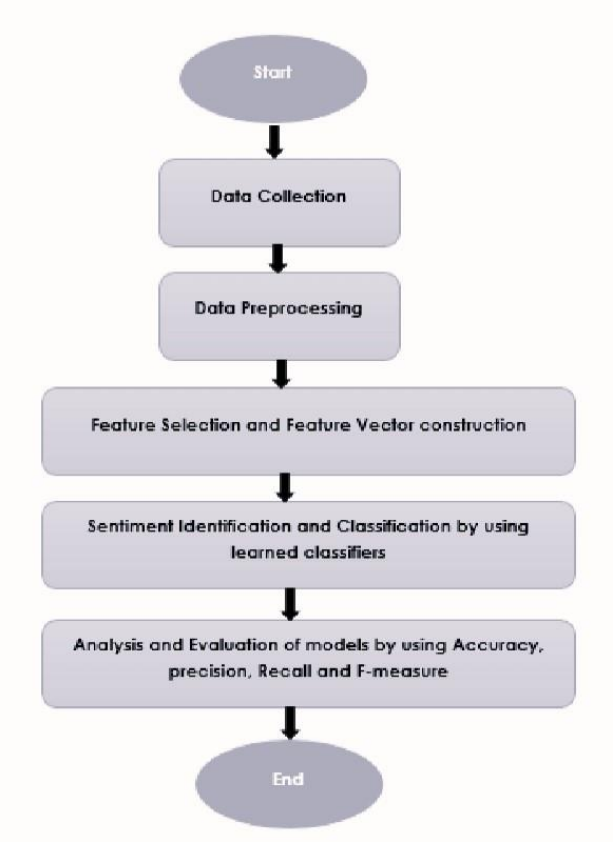

Fig. 1. Flow chart

C. *BLOCK DIAGRAM*

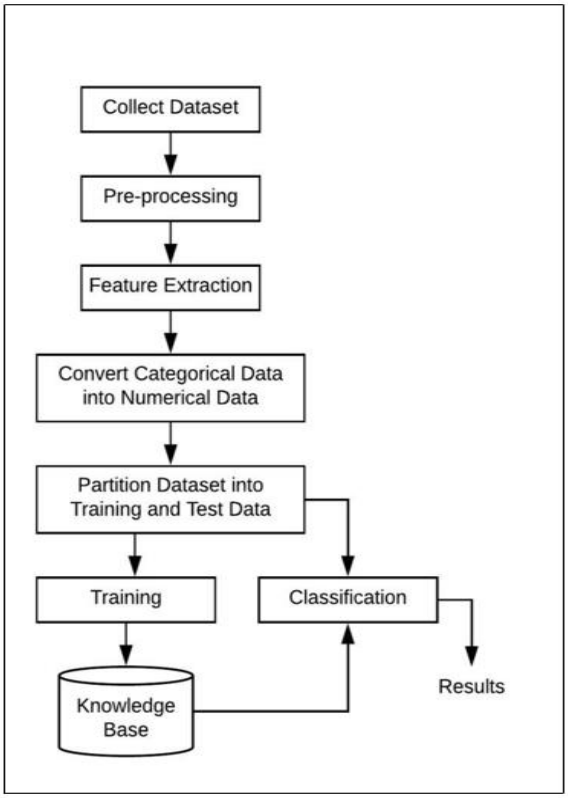

Fig. 2. Block Diagram

IV. PROCESSING TECHNIQUES

A. *Data Collection:*

Gather a comprehensive dataset that includes relevant variables related to happiness from the World Happiness Report and other reputable sources. Ensure the dataset covers a diverse range of countries, demographics, and happiness indicators to capture a representative sample.
https://1drv.ms/u/s!AvCypxdKGfSGgxDyQuwC0-Eyry7Q

B. *Data Preparation:*

Handle missing values, eliminate duplicates, and deal with outliers to clean up the dataset. Perform necessary data transformations, such as normalization or feature scaling, to ensure consistency and comparability across variables.

	Country name	Ladder score	Standard error of ladder score	upperwhisker	lowerwhisker	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Ladder score in Dystopia	Explained by: Log GDP per capita
0	Finland	7.804000	0.036000	7.875000	7.733000	10.792000	0.969000	71.150000	0.961000	-0.019000	0.182000	1.778000	1.888000
1	Denmark	7.586000	0.041000	7.667000	7.506000	10.962000	0.954000	71.250000	0.934000	0.134000	0.196000	1.778000	1.949000
2	Iceland	7.530000	0.049000	7.625000	7.434000	10.896000	0.983000	72.050000	0.936000	0.211000	0.668000	1.778000	1.926000
3	Israel	7.473000	0.032000	7.535000	7.411000	10.639000	0.943000	72.697000	0.809000	-0.023000	0.708000	1.778000	1.833000
4	Netherlands	7.403000	0.029000	7.460000	7.346000	10.942000	0.930000	71.550000	0.887000	0.213000	0.379000	1.778000	1.942000
5	Sweden	7.395000	0.037000	7.468000	7.322000	10.883000	0.939000	72.150000	0.948000	0.165000	0.202000	1.778000	1.921000
6	Norway	7.315000	0.044000	7.402000	7.229000	11.088000	0.943000	71.500000	0.947000	0.141000	0.283000	1.778000	1.994000
7	Switzerland	7.240000	0.043000	7.324000	7.156000	11.164000	0.920000	72.900000	0.891000	0.027000	0.266000	1.778000	2.022000
8	Luxembourg	7.228000	0.069000	7.363000	7.093000	11.660000	0.879000	71.675000	0.915000	0.024000	0.345000	1.778000	2.200000
9	New Zealand	7.123000	0.038000	7.198000	7.048000	10.662000	0.952000	70.350000	0.887000	0.175000	0.271000	1.778000	1.842000
10	Austria	7.097000	0.040000	7.176000	7.018000	10.899000	0.888000	71.150000	0.855000	0.102000	0.497000	1.778000	1.927000
11	Australia	7.095000	0.044000	7.180000	7.009000	10.821000	0.934000	71.050000	0.890000	0.198000	0.496000	1.778000	1.899000
12	Canada	6.961000	0.042000	7.042000	6.879000	10.773000	0.929000	71.400000	0.874000	0.153000	0.420000	1.778000	1.881000
13	Ireland	6.911000	0.044000	6.996000	6.825000	11.527000	0.905000	71.300000	0.874000	0.092000	0.358000	1.778000	2.152000
14	United	6.894000	0.047000	6.986000	6.802000	11.048000	0.919000	65.850000	0.800000	0.137000	0.689000	1.778000	1.980000

Fig. 3. Preprocessed data

C. *Data Visualization:*

The pre-processing of data in order to visualize the World Happiness Report. Among them are cleaning techniques for outliers, inconsistent data, and missing values. For missing values, imputation techniques can be used, and outliers can be corrected. Scaling, also known as normalization, guarantees equitable comparisons across variables with disparate scales. For compatibility, categorical variables might need to be converted to numerical data as the ladder score is in numeric. To produce precise and insightful visualizations of the various happiness indicators included in the World Happiness Report.

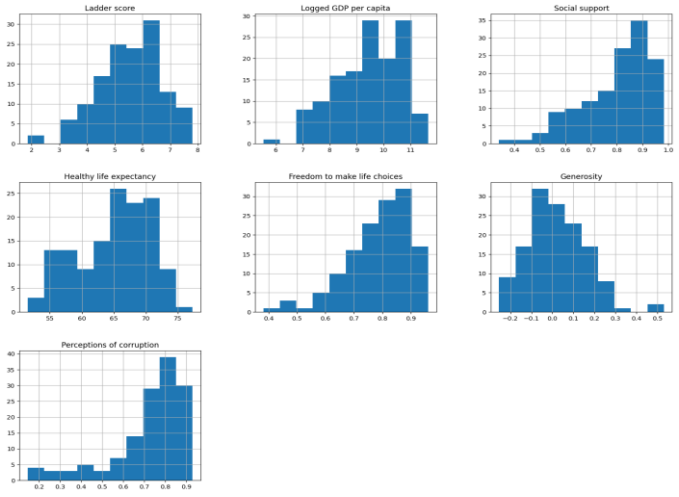

Fig. 4. Univariate analysis visualizations (Histogram)

D. *Feature Engineering:*

Extract and select meaningful features from the dataset that influence happiness scores. This may include economic indicators, social factors, environmental variables, and cultural dimensions. Consider incorporating additional derived features or aggregating indicators to enhance the predictive power of the model.

	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
0	1.170663	0.020420	1.142420	0.022928	-0.267674	-3.082101
1	1.315511	1.240957	1.160436	1.355912	0.927328	-3.002695
2	1.259276	0.010648	1.304562	1.374097	1.528734	-0.325573
3	1.040299	1.154298	1.421124	0.219375	-0.298916	-0.098699
4	1.298470	1.051882	1.214483	0.928574	1.544354	-1.964743

Fig. 5. Selected Attributes after preprocessing

E. *Model Selection:*

Experiment with various machine learning algorithms suitable for regression tasks, such as linear regression, random forests, gradient boosting, or ridge and lasso. Explore different algorithms to identify the one that yields the best performance in predicting happiness scores.

F. *Training and Evaluation:*

Train the chosen models using the training data after dividing the dataset into training and testing datasets using train_test_split function. Assess the models' ability to predict happiness scores by utilizing suitable assessment metrics, such as R-squared or mean squared error. Cross-validation can be used to verify the robustness of the model.

```
# train test split
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_state=0)
```

Fig. 6. Dividing the dataset into training and testing datasets using train_test_split function

```
# training the model
model = model.fit(X_train,y_train)
rf.fit(X_train, y_train)
```

```
RandomForestRegressor
RandomForestRegressor(random_state=42)
```

Fig. 7. Training the data set with random forest regressor

```
# testing the model
pred = model.predict(X_test)
pred_rf= rf.predict(X_test)
```

Fig. 8. Testing the data set with random forest regressor

G. *Hyperparameter Tuning:*

To maximize the performance of the selected models, adjust their hyperparameters. Use methods such as grid search or random search to investigate various hyperparameter combinations and find the best setup.

H. *Comparative Analysis:*

Determine which model predicts happiness scores the best by comparing its performance to that of the others. Examine each model's advantages and disadvantages to learn more about how appropriate and comprehensible it is.

	Regression Models	R2_Scores	Mean_Squared_error	Mean_Absolute_error
0	random forst	0.761707	0.184987	0.345507
1	gradient boosting	0.691153	0.239758	0.386737
2	ridge	0.673238	0.253665	0.384666
3	lasso	-0.057502	0.820939	0.384666

Fig. 9. Comparison of the models

I. Testing on Unseen Data:

Evaluate the selected model on unseen data to assess its generalization ability. This step helps ensure that the model can accurately predict happiness scores for new countries or observations outside the training dataset.

J. Validation and Real-world Application:

Validate the model's performance by applying it to real-world scenarios or deploying it in a controlled environment. Monitor the model's predictions and iterate on improvements if necessary. For example, like a user-friendly website.

V. RESULT AND FINDINGS

In the analysis of the World Happiness Report data, the effectiveness of several regression models, such as Random Forest, Gradient Boosting, Ridge, and Lasso Regressors, was assessed. The R2 score, a metric that measures the proportion of variance in the dependent variable (ladder scores) that is predictable from the independent factors.

It's interesting to note that out of all the evaluated algorithms, the Random Forest Regressor proved to be the most accurate model. In the test dataset, it performed better than the Gradient Boosting, Ridge, and Lasso Regressors in terms of R2 score. The Random Forest model appears to have fit the underlying patterns in the World Happiness Report data better during the testing phase, based on the higher R2 score.

The Random Forest algorithm's ensemble nature, which integrates several decision trees to reduce overfitting and improve prediction accuracy, may be the reason for its better performance. Because of its adaptability, the Random Forest model is especially well-suited for forecasting ladder scores, since it can capture relationships within the data. Although Ridge and Lasso Regressors and Gradient Boosting Regressors are strong algorithms on their own, the unique properties of the World Happiness Report data may make the Random Forest method more advantageous. In order to ensure an accurate representation of happiness rankings and contributing factors within the framework of the World Happiness Report, this result highlights the importance of methodically comparing and choosing regression models based on their performance metrics.

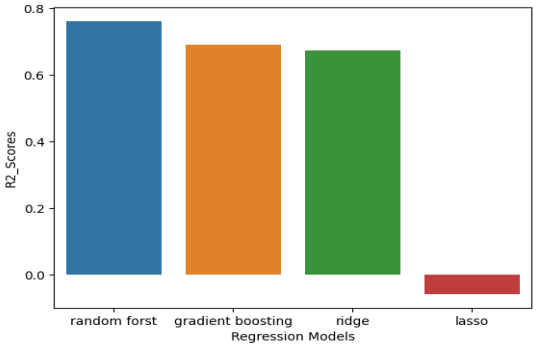

Fig. 10. R2score on test data of various models

VI. FUTURE SCOPE

Website Development: Through the development of a user-friendly website and the integration of data from the World Happiness Report, the project will offer users the opportunity to input their country's data and personal information, resulting in a predicted happiness score. This allows individuals to obtain a better comprehension of the elements that go into happiness and provides a platform for exploring potential areas for improvement.

VII. CONCLUSION

In conclusion, the World Happiness Report project provides valuable insights into the factors influencing happiness at a global and country level. By leveraging the comprehensive analysis and rankings provided in the report, the project aims to predict happiness scores based on user inputs, empowering individuals to understand and explore the happiness levels of their respective countries.

REFERENCES

[1] Andrew N. Rowan, WellBeing International(WellBeing News) ,“World Happiness Report 2023”.

[2] Charles Alba,”A Data Analysis Of The World Happiness Index And Its Relation To The North-South Divide”, Pennsylvania State University - Main Campus.

[3] Francesco Sarracino HSE-LCSR; STATEC, Kelsey O'Connor,STATEC Research; IZA Institute of Labor Economics, “A Measure of Well-Being Efficiency Based on the World Happiness Report, IZA Discussion Paper No. 15669” .

[4] Francesco Sarracino, Kelsey O'Connor, “analysing happiness index as a measure along with its parameters and strategies for improving india’s rank in world happiness report”.

[5] John Helliwell, Richard Layard and Jeffrey Sachs” World happiness report(2012)” . The Earth Institute, Columbia University, New York, USA.

[6] John Helliwell, Richard Layard and Jeffrey Sachs, “ world happiness report (2017)”.

[7] M. Mujiya Ulkhaq, Department of Economics and Management, University of Brescia, Brescia 25121, Italy, Arga Adyatama PT Algoritma Data Indonesia, South Jakarta 12950, Indonesia,”Clustering countries according to the world happiness report 2019”.

[8] Moaiad Ahmad Khder, Department of Computer Science – Applied Science University-Bahrain. Mohammad Adnan Sayfi, Data Engineer – Umniah Jordan. Samah Wael Fujo Data Scientist – Nasser Artificial Intelligence Research and Development Centre - Nasser Vocational Training Center-Bahrain ,“Analysis of World Happiness Report Dataset Using Machine Learning Approaches “.

[9] Michael Kaufman, Andrew M. Guest, “What the World Happiness Report doesn’t see: The sociocultural contours of wellbeing in northern Tanzania”.

[10] Sarah Ahtesham, “analysing happiness index as a measure along with its parameters and strategies for improving india’s rank in world happiness report”.